Data augmentation and image understanding

PhD Thesis
Institute of Cognitive Science
University of Osnabrück

Alex Hernández-García

Advisor: Prof. Peter König

July 2020

Doctoral committee:
Jeffrey Bowers
Konrad P. Kording
Graham W. Taylor
Peter König (advisor)
Gordon Pipa (chair)

Foreword

There are certainly many people I am grateful to for their support and love. I hope I show them and they feel my gratitude and love often enough. Instead of naming them here, I will make sure to thank them personally. I am also grateful to my PhD adviser and collaborators. I acknowledge their specific contributions at the beginning of each chapter. I wish to use this space for putting down in words some thoughts that have been recurrently present throughout my PhD and are especially strong now.

The main reason why I have been able to write a PhD thesis is *luck*. Specifically, I have been very lucky to be a privileged person. In my opinion, we often overestimate our own effort, skills and hard work as the main factors for professional success. I would not have got here without my privileged position, in the first place.

There are some obvious comparisons. To name a few: During the first months of my PhD, Syrians had undergone several years of war[1] and citizens of Aleppo were suffering one of the most cruel sieges ever[2]. Hundreds of thousands were killed and millions had to flee their land and seek refuge. While I was visiting Palestine after a PhD-related workshop in Jerusalem, 17 Palestinian people were killed by the Israeli army in Gaza during the Land Day[3] protests. Almost 200 hundred more human beings were killed and thousands were injured by snipers in the weeks thereafter[4]. Also during my PhD, hundreds of thousands of Rohingya[5] nationals of Myanmar (formerly Burma) were forced to flee their country due to ethnic and religious persecution. Many people were killed, tortured and thousands of women and girls were raped. An enormous fraction of the population in Africa live below the poverty line, threatened by starvation, conflict, diseases and climate change. Many are forced to cross the continent to seek a safer life in Europe, going through one of the most dangerous migration routes in the world. More than 12,000 human beings have lost their lives in the Mediterranean Sea[6] and many more are thought to have died in the Sahara Desert, only since I started my PhD. Around the world, thousands of women are killed by men every year and many live under extremely unsafe conditions[7]. People in these circumstances, regardless of how intelligent and hard-working, can hardly consider the idea of pursuing a PhD. These are only a few well-known, extreme situations around the globe. The list would go much longer, highlighting the privilege of those who are born at the lucky side of the world. I have also been lucky in subtler ways.

First of all, I am a man, and my skin is white. I grew up in Madrid, the capital of a member state of the European Union. This is already privileged with respect to most of the rest of the world, and to most of the rest of my country, Spain. Even though I was not born to a rich family, my parents went to university and became a nurse and a teacher. They always encouraged me and my siblings to study, while giving us the freedom to choose what we wanted. They showed us the pleasure of reading and learning. Not all children, even in the privileged areas of the privileged countries, grow in such positive environments. I have also been lucky that I have not had to go through a separation of my parents. I found supportive friends and had most of the rest of my family close by during my childhood. I worked in the summers to earn some pocket money and certain independence, but I was lucky enough to be able to focus on my studies during the year. Some of my friends could not afford to go to uni. My parents could also afford to cover a year abroad as an Erasmus student. Nothing of this was a merit of mine, but pure *luck*. Without all these positive initial conditions I may not have gained sufficient intellectual and emotional development. Of course I did study and work hard, but many work a lot harder just to survive.

Knowing of my privileged position encourages me to keep the focus and try to make the best out of my time and work. I strongly believe that our everyday actions make a change in the world and that the progress of science will lead us to a fairer planet. This dissertation will make little change, if anything, but the collective effort of many has the potential to make a big impact. May this keep on steering my motivation.

July 2020

[1] https://en.wikipedia.org/wiki/Syrian_civil_war#Timeline
[2] https://en.wikipedia.org/wiki/Battle_of_Aleppo_(2012-2016)
[3] https://en.wikipedia.org/wiki/Land_Day
[4] https://en.wikipedia.org/wiki/2018-19_Gaza_border_protests
[5] https://en.wikipedia.org/wiki/Rohingya_refugees_in_Bangladesh
[6] https://data2.unhcr.org/en/situations/mediterranean
[7] https://en.wikipedia.org/wiki/Violence_against_women

Abstract

Interdisciplinary research is often at the core of scientific progress. As an example, artificial neural networks, currently an essential tool in many applications that require learning from data, were originally inspired by insights from biological neural networks. Since its inception as a field, the progress of artificial intelligence has at times converged and at times diverged from the field of neuroscience. While occasional divergence can be fruitful, in this dissertation we will explore some advantageous synergies between machine learning, cognitive science and neuroscience.

In particular, this thesis focuses on vision and images. The human visual system has been widely studied from both behavioural and neuroscientific points of view, as vision is the dominant sense of most people. In turn, machine vision has also been an active area of research, currently dominated by the use of artificial neural networks. Despite their origin and some similarities with biological networks, the recent progress in neural networks for image understanding has shown signs of divergence from neuroscience. One likely cause is the focus on benchmark performance, regardless of *what* the models learn. This work focuses instead on *learning representations* that are more aligned with visual perception and the biological vision. For that purpose, I have studied tools and aspects from cognitive science and computational neuroscience, and attempted to incorporate them into machine learning models of vision.

A central subject of this dissertation is data augmentation, a commonly used technique for training artificial neural networks to *augment* the size of data sets through transformations of the images. Although often overlooked, data augmentation implements transformations that are perceptually plausible, since they correspond to the transformations we see in our visual world—changes in viewpoint or illumination, for instance. Furthermore, neuroscientists have found that the brain invariantly represents objects under these transformations. Throughout this dissertation, I use these insights to analyse data augmentation as a particularly useful inductive bias, a more effective regularisation method for artificial neural networks, and as the framework to analyse and improve the invariance of vision models to perceptually plausible transformations. Overall, this work aims to shed more light on the properties of data augmentation and demonstrate the potential of interdisciplinary research.

The code produced in this thesis is open source and available at `www.github.com/alexhernandezgarcia`

Contents

Chapter 1

Introduction

Visual information plays a remarkably prominent role in how humans and many other animals perceive the world. By way of illustration, about 27 % of the cerebral cortex of humans and 52 % of the macaque monkey's is devoted to vision (Van Essen, 2003). This involves billions of neurons and connections which give rise to very complex and sophisticated mechanisms, such as visual attention and object recognition. Understanding the way visual information is processed in the brain and how it affects our behaviour is a very active area of research in cognitive science and neuroscience. In turn, developing artificial algorithms that mimic some aspects of biological vision by learning patterns from collections of image data is another active and currently fruitful area of research in machine learning and computer vision. While all these disciplines are rooted in significantly different origins and make use of distinct tools, there are grounds to defend that the interdisciplinary study of these areas of research can highly benefit each other (Bengio et al., 2015; Marblestone et al., 2016; Hassabis et al., 2017; Bowers, 2017; Richards et al., 2019; Kietzmann et al., 2019; Lindsay, 2020; Saxe et al., 2020).

In this thesis, we study several aspects of vision and image understanding, such as visual object recognition and visual attention, using the methods and techniques of various disciplines. In particular, we approach machine visual object recognition through deep artificial neural networks, compare some of their properties with the visual cortex, and draw inspiration from visual perception and biological vision to improve the artificial models; we employ eye tracking to analyse the global salience of images when humans look at competing stimuli; and we study the effectiveness of visual salience to identify images from fMRI data. Overall, this dissertation aims at integrating knowledge and methodologies of various disciplines to further advance our understanding of biological and artificial vision.

Summarised, the specific contributions of this thesis are the following: we shed light on the heavily understudied role of data augmentation—transformations of the input data—on training artificial neural networks for image object recognition, and show that it is more effective than the most popular regularisation techniques. Additionally, we found that models trained with heavier image transformations may learn internal representations more aligned with the activations measured in the visual cortex of the brain. Further, we demonstrate how data augmentation can be used to incorporate inductive biases—useful priors—from visual perception and biological vision, in particular the invariance to identity-preserving transformations. Separately, we propose the concept of image global salience, a property of natural images that reflects the likelihood of a stimulus to attract the gaze of a human observer when presented in competition with other stimuli. Finally, we studied the correlation of image salience with the measured activations in the visual cortex.

The interdisciplinary nature of this thesis is rooted in the breadth of my[1] personal interests as

[1]Except in the cases where I refer explicitly to me—the author of this thesis—as an individual or intend to express my

well as in the opportunities that the characteristics of my PhD programme offered to me. Having a background on computer science and electrical engineering, already my Bachelor's thesis (Hernández-García, 2014) addressed the problem of predicting subjective perception from audiovisual content, work continued in my Master's thesis, where I used psychophysical measurements (Hernández-García et al., 2017). My journey towards cognitive science went on when I was admitted as a PhD candidate in the Institute of Cognitive Science of the University of Osnabrück, with Professor Peter König—although in practice I lived and worked in Berlin. The PhD programme also gave me the opportunity to get in touch, for the first time, with neuroscience, as I became a fellow of the Marie Skłodowska-Curie Innovative Training Network "Training the next generation of European visual neuroscientists" (NextGenVis[2]). In spite of the challenges of being an outlier in the cohort of PhD students as a machine learning scientist, the breadth of my interests certainly expanded. As a mandatory aspect of the Marie Skłodowska-Curie grant, I had to carry out two 2-months internships at laboratories of the partnership or related. This gave me the opportunity to work at the Spinoza Centre for Neuroimaging in Amsterdam with Dr. Serge Dumoulin, where I worked on the project that is described in Chapter 8; and at the Cognition and Brain Sciences Unit of the University of Cambridge with Dr. Tim Kietzmann, where I started the work that yielded Chapter 6. I am very greateful to Serge, Tim and Peter for these opportunities.

The breadth of one's interests and work is certainly at odds with depth. However, there is probably not a single sweet spot in the trade-off between breadth and depth valid for all scientists. I understand science as a collaborative ecosystem that is most effective if scientists are distributed across the whole spectrum of the breadth-depth dichotomy. This thesis—and my work so far—is an attempt to explore the interdisciplinary approach to science, in particular to the study of learning systems—artificial and biological—specialised in visual information.

1.1 Learning to see

Vision has evolved as one of the most advanced perceptual mechanisms in humans and many other animals, and is the main source of information for most of us to navigate and understand the world around us. Among the multiple high-level perceptual abilities achieved by our visual system, one of the most sophisticated and best studied is visual object recognition: under an enormous variety of conditions, humans can easily distinguish objects of different categories as well as identify individual objects within the same class (Logothetis & Sheinberg, 1996). Being such an important aspect of human perception, automatically finding and identifying objects from digital images has been as well an active area of research in computer vision and a technological challenge for several decades.

The task of image object recognition in computer vision can be defined in its simplest case as the categorisation of an image into one class from a set of pre-defined object classes, according to the main object present in the image by extracting and processing the visual information, that is the pixel values. Object recognition is closely related to other tasks such image object detection (Zhao et al., 2019) and content-based image retrieval (Latif et al., 2019; Zhou et al., 2017). The generalisation of this broader set of of tasks is known as *image understanding*. Ultimately, the goal of artificial image understanding is to extract rich and complex information from a digital image, as close as possible to biological vision and visual perception.

During several decades, object recognition and the related tasks of image understanding were predominantly approached by extracting handcrafted features from the images to then train machine learning classifiers. As a result, much of the computer vision literature was dominated by the proposal and refinement of image processing techniques, such as edge detectors, line and curve detectors like the Hough transform (Duda & Hart, 1972), scale-invariant feature descriptors such like SIFT (Lowe, 2004) and HOG (Dalal & Triggs, 2005) or Haar-like features (Lienhart &

personal opinion, I will use, in general, the plural form of the first person in order to acknowledge the contribution of some collaborators to certain parts of the work and keep a consistency throughout the whole thesis.

[2]www.nextgenvis.eu

Maydt, 2002), among many others. These features are usually combined using bag-of-visual-words techniques (Sivic & Zisserman, 2003) to train classifiers such as support vector machines (SVM) (Cortes & Vapnik, 1995); or more sophisticated approaches such as Fisher kernels (Perronnin & Dance, 2007). Today, these techniques are informally known as *traditional* machine learning or *traditional* computer vision.

What currently is considered *modern* machine learning are *deep artificial neural networks* (ANN). A major breakthrough in computer image object recognition occurred in 2012 when Krizhevsky et al. (2012) presented results of a neural network algorithm that nearly halved the previous error measure on the large-scale benchmark data set ImageNet (Russakovsky et al., 2015). This attracted an increasing amount of attention towards this kind of algorithms, leading to a rapid development of the subfield of machine learning known as *deep learning*.

ANNs are, however, not exactly *modern*. The history of neural networks in the scientific literature can be traced back at least to the 1940s, when the first theories of biological learning were proposed (McCulloch & Pitts, 1943; Hebb, 1949). Biological learning and biological neural networks inspired the development of artificial models such as the perceptron (Rosenblatt, 1958) and later the Neocognitron (Fukushima & Miyake, 1982)—a precursor of current convolutional neural networks—directly inspired by the findings by Hubel & Wiesel (1959) about the receptive fields of neurons in the cat's visual cortex. ANNs gained some popularity in the 1980s and 1990s with the development and application of the *backpropagation* principle (Rumelhart et al., 1986) and other successful proposals, such as the long-short term memory (LSTM) recurrent neural networks (Hochreiter & Schmidhuber, 1997). Nonetheless, the progress and adoption of neural networks until 2012 was slow and minor, overshadowed by other algorithms such as the SVM.

One of the main differences—and arguably, advantages—of deep learning with respect to traditional machine learning algorithms is that they are able to extract relevant features for a certain task, such as image object recognition, more directly from the data. This allows to apply similar learning principles, techniques and architectures to a more general set of data modalities and tasks. This can be regarded as a step towards finding a general learning principle, a hypothesis suggested in biological learning, upon the observation of the highly regular structure found in the neocortex across different brain areas and species (Douglas et al., 1989; Harris & Shepherd, 2015). This *philosophy* of letting an algorithm learn the relevant features—*representation learning*—from the available data can be regarded as the other end of *symbolic artificial intelligence* (AI) (Haugeland, 1989), whose approach was to programme an agent to perform certain tasks by directly manipulating symbols and manually implementing rules to simulate some behaviour. The traditional machine learning and computer vision algorithms used for many years would be somewhere in between symbolic AI and the philosophy of deep learning.

Despite the significant practical and conceptual step forward originated by the recent progress and adoption of deep learning, the claims or beliefs that ANNs are or will be able to automatically learn anything from data without human intervention are overstated. Naturally, artificial neural networks are also subject to the *no free lunch theorem*: no learning algorithm is better than any other at classifying unobserved data points, when averaged over all possible data distributions (Wolpert, 1996). In other words, in order to learn anything about the world it is necessary to introduce some *priors* or *inductive biases* about the world.

In this light, although it is often claimed that deep learning is a shift away from the traditional approach of hand-engineering features to "automatically" learning them with "a general-purpose learning procedure" (LeCun et al., 2015), it can also be seen as a shift of inductive biases. The priors used by deep learning are definitely *more* general-purpose than those used in traditional machine learning, but neural networks do not remove the need for inductive biases—and they will never do. While a few years ago we handcrafted visual descriptors that could discriminate object classes in a set of images, we now handcraft types of layers, architectures, parameter initialisation strategies, regularisers and optimisation methods, to name a few elements of deep learning[3]. The

[3]Even approaches such as neural architecture search (Zoph & Le, 2016), which aim at further automating machine learning, not only use a tremendous amount of computational resources—with a proportional environmental impact—but their efficacy has been questioned, as reviewed by Gencoglu et al. (2019), among others.

current success of deep learning is the result of a collective effort in exploring the combinations of these elements that work best for different tasks, presented in a large body of scientific literature.

We argue that the new landscape opened by deep learning is undoubtedly a significant step towards a more natural and general way of processing data, especially because it allows to train learning algorithms almost end-to-end from nearly naturalistic sensory signals, such as digital images or sound (Saxe et al., 2020). However, we should not neglect the need for searching better inductive biases, that is incorporating prior information about the tasks we want to solve, since without it learning would be much less efficient or not possible at all. We know this from statistical learning theory, but also from the innate genetic inductive biases provided by evolution in nature, which seem to predispose organisms to quickly learn and adapt to their specific environment (Zador, 2019). Hence, studying biological learning systems seems like a natural approach to draw inspiration for improving artificial learning algorithms (Hassabis et al., 2017; Nayebi & Ganguli, 2017; Lindsay & Miller, 2018; Malhotra et al., 2020).

A key pair of ingredients in the development and success of deep learning was the increase in available computational power, alongside the publication of large data sets. The comparably much smaller data sets and reduced computational resources several decades ago has been argued to be one of the reasons why researchers could not prove the capabilities of ANNs until recently. Compared to other machine learning models, neural networks excel at learning from large data sets, but in some cases require much and specialised computer power, like graphical processing units (GPU). In computer vision specifically, the efforts in hand-labelling large data sets in the last decades set the grounds for deep learning to unleash its potential. Some of these data sets, which we have used in the experiments presented in this thesis, are MNIST, 70,000 small greyscale images of digits (LeCun et al., 1998); CIFAR-10 and CIFAR-100, 60,000 small colour images of objects from 10 and 100 classes, respectively (Krizhevsky & Hinton, 2009); and ImageNet (ILSVRC 2012), 1.3 million high-resolution natural images of 1,000 object classes (Russakovsky et al., 2015).

With the availability of these benchmark data sets to train and compare different algorithms, most of the research efforts in the last years have been used to improve different aspects of the neural network architectures and the training process: parameter initialisation strategies (Glorot & Bengio, 2010; He et al., 2015), activation functions (Glorot et al., 2011), normalisation layers (Ioffe & Szegedy, 2015), stochastic optimisation methods (Duchi et al., 2011; Kingma & Ba, 2014), network architectures (Springenberg et al., 2014; Simonyan & Zisserman, 2014; He et al., 2016; Huang et al., 2017), and so on. The collaborative effort of many researchers has been essential to develop these and many other methods that have advanced the performance and our understanding of deep neural networks. Meanwhile, beyond the creation and publication of data sets, little attention has been paid to the data, which was considered fixed and given, since the goal is generally to improve the state of the art results on the common benchmark data sets. However, we have noted that the availability of data was key for the success of deep learning in image object recognition.

The need for larger amounts of labelled examples to train neural networks led some researchers to create data sets, and also led many to think of ways of easily create new examples. A straightforward way to extend a data set, especially in the image domain, is through *data augmentation*. Data augmentation, broadly defined, consists of synthetically expanding a data set by applying transformations on the available examples that preserve the ground truth labels. Data augmentation has been used in image object recognition at least since the 1990s (Abu-Mostafa, 1990; Simard et al., 1992); has been identified as critical component of many successful models, such as AlexNet (Krizhevsky et al., 2012); and is ubiquitous in both deep learning research and application. Nonetheless, despite its popularity, data augmentation has surprisingly remained a largely understudied component of machine learning: many in the field use it, know that it helps generalisation and intuitively understand why, but little is known about its interaction with other techniques or whether its actual potential goes beyond simply increasing the number of training examples.

A significant part of this thesis aims at filling this gap and shedding light on the role of data augmentation for visual object recognition with deep neural networks. We argue that data aug-

mentation has been heavily understudied because it has been looked down on by the machine learning community, regarded almost as a *cheating* technique, rather than as a method that deserves analysis and attention. In this dissertation, we contend that data augmentation is interesting beyond simply providing new examples. We analyse data augmentation as regularisation, but draw significant differences with respect to classical regularisation. We also analyse the potential of data augmentation for training models that learn robust visual representations, taking inspiration from properties of the visual cortex. Overall, we here aim to explore the connections between data augmentation, visual perception and biological vision.

1.2 Why data augmentation?

Having a background in image processing and having carried out research projects with traditional, handcrafted visual descriptors (Hernández-García et al., 2017), when I first started training neural networks with image data at the beginning of my PhD, it was the most natural approach for me to apply transformations on the images in my data sets to get some more data points *for free*. As I was learning about deep learning, I became increasingly interested in the generalisation properties of neural networks and in regularisation methods. In my toy experiments, I noticed that while the most popular regularisation methods, that is weight decay and dropout, did improve the test performance of the models, the true boost in generalisation seemed to be provided by the image transformations that I had coded, that is data augmentation. This made sense to me: we know from statistical learning theory that generalisation can improve by finding the right complexity of the model, that is by accurately tuning the regularisation; but generalisation should always improve with more training examples (see Section 2.1.2).

I became even more intrigued about these observations and got additional insights after reading "Understanding deep learning requires rethinking regularization", by Zhang et al. (2017a). This article, among other results, included the performance of a few models trained with and without weight decay, dropout and data augmentation. The superiority of data augmentation was also apparent by carefully analysing the results in the tables, but this fact is nearly ignored in the paper, since all three methods—weight decay, dropout and data augmentation—are considered the same type of regularisation (see Chapter 4). I had the chance to chat with Dr. Chiyuan Zhang at his poster presentation at the International Conference on Learning Representations in 2017. While he was arguing how the generalisation bounds based on the Rademacher complexity (see Section 2.1.2) may not explain the generalisation performance of deep neural networks, I asked what would the n—the number of training examples—be in the formula if they trained with data augmentation. Of course there is no straightforward answer to that question, so that got me thinking.

Since data augmentation was so remarkably effective in improving the test performance of neural networks, I assumed there should be literature on the topic that explained how exactly data augmentation affects generalisation and empirical comparisons with other regularisation methods. Unfortunately, I could not find much, except than corroborating that everybody seemed to use data augmentation in their papers to push the performance of their models towards the state of the art (Krizhevsky et al., 2012; Springenberg et al., 2014). I could think of three types of reasons for this lack of literature on data augmentation: One, "it is *obvious* that data augmentation is beneficial". Two, "it is *too complicated* to study". Three, "it is *not interesting*". The first reason was not very satisfying, scientifically. The second one was actually a reason for me to try to shed some light. While it is probably a mix of all, I think the third point was the actual reason why data augmentation was disregarded: As mentioned in Section 1.1, the philosophy of the re-emerging deep learning field was to let a model learn "good features ... automatically using a general-purpose learning procedure" (LeCun et al., 2015). Handcrafting some image transformations to augment a data set seemed closer to the *old-fashioned* traditional machine learning methods and against the deep learning philosophy. As a matter of fact, data augmentation seemed to be considered *cheating* and many papers would include results of their new method or architecture *without* data augmentation to show that it actually worked (Goodfellow et al., 2013; Graham, 2014; Larsson

et al., 2016)—this ablation studies were however not carried out with other techniques, such as weight decay or dropout.

The fact that data augmentation increases the number of training examples, even though breaking the assumption of independent and identically distributed sampling, is only the most obvious advantage: the tip of the iceberg. In what follows, we will outline the interpretation of data augmentation as a remarkably useful inductive bias directly connected to visual perception, and make explicit some connections with properties of the visual cortex.

In the application of deep learning we have considered so far, image understanding and, in particular, object recognition, it is easy to get stuck in the specific goal of the task at hand that is often performed: to obtain a high classification performance in the test set of the benchmark data set. However, it is always useful to take a step back to see the forest for the trees. The objective for, at least, the research community is not to incrementally improve the state of the art performance on the benchmark data sets, but to truly develop good models of image understanding. Furthermore, when we talk about object recognition, in general we do not mean object recognition for any arbitrary visual system or modality of the many in nature, but we mean *human* object recognition, that is recognising the object classes that are relevant for humans, from photos or videos that actually resemble how we perceive the world, that is natural images[4].

Since we are interested in human object recognition, we argue that we should always keep an eye on what we know from visual perception and, ideally, biological vision too. The transformations that are most commonly applied in data augmentation schemes are *perceptually plausible*: the resulting image preserves the properties of the perceived visual world as well as, in the case of object recognition, the object class[5]. Especially in the image domain, it is straightforward to identify a large number of perceptually plausible transformations—equivariant transformations, some colour adjustments, blurring, etc. Some examples are shown in Figure 1.1—and it is well-known since decades ago how to apply them to digital images (Gonzalez & Woods, 2018). Hence, the combination of having tools for performing the transformations and expert domain knowledge—visual perception—provides us with a large-capacity generator of new examples and an approximate *oracle* of the target function, for a relevant subset of the input space.

The access to an oracle of the target function serves as a remarkably effective inductive bias, which is at the very essence of machine learning. Data augmentation exploits this oracle, constructed from our knowledge about perception to densely populate the relevant regions of the high-dimensional input space. That is, the different views of the objects that are perceptually plausible. Richards et al. (2019) argue that in order to train neural networks "the three components specified by design are the objective functions, the learning rules and the architectures". Throughout this dissertation we will argue that incorporating inductive biases through the data, possibly in combination with the objective function (Chapter 6), is another effective way to learn better, more robust representations.

Figure 1.1: Some examples of perceptually plausible image transformations applied on the same image, from ImageNet.

[4]Some particular subfields of computer vision are devoted to non-human vision, for example multispectral imagery (Audebert et al., 2019).

[5]Other types of transformations in which perceptual plausibility is not preserved have been successfully used in computer vision. One popular example is *mixup* (Zhang et al., 2017b), which performs the weighted average of the pixels of two images—and their labels. Another example is data augmentation in feature space (DeVries & Taylor, 2017). However, in this thesis we are interested in exploring connections between machine learning and visual perception and thus we will consider only perceptually plausible image transformations.

1.3 Rethinking supervised learning

The re-emergence of deep learning during the last decade due to the noteworthy achievements of artificial neural networks built up a sort of philosophy that anything could be automatically learnt from data without human intervention, in contrast to the previous approach that required, indeed, a higher degree of manual design. However, this promise is misleading, since it misses the fact that the success of deep networks has been partly due to the human effort of manually labelling thousands of images and other data modalities (Russakovsky et al., 2015; Cao et al., 2018). As a matter of fact, the need for much data has been at the core of some strong criticisms of deep learning (Marcus, 2018). We argue instead that neither is deep learning some sort of exceptional solution to learn without inductive biases, nor is it a hopeless model class because it requires large data sets. On the contrary, we contend that the competitive advantage of artificial neural networks is that they are deep universal function approximators (Hornik, 1991) that have been proven to *be able to* learn from large collections of data. While other models are known to scale poorly as the amount of training data increases, ANNs excel at fitting the training data and applying interpolation on unseen examples (Belkin et al., 2019; Hasson et al., 2020). This is a feature, not a bug. We will make faster progress if we exploit the advantages of deep learning, while being aware of its limitations.

Data augmentation, as we have seen, is an effective way of using this property: it generates examples in the regions of the input space where the model should learn how to interpolate. A commonly seen argument to defend that this should not be necessary and neural networks should be able to generalise from few examples is that humans and other animals learn—to categorise objects, for example—with very little supervision (Vinyals et al., 2016; Marcus, 2018; Morgenstern et al., 2019; Zador, 2019). In this section, we will elaborate on our views on why we think this may be a misconception that can lead us astray and how we can benefit from rethinking the notions of supervised and unsupervised learning. We will do so by taking a look into learning theory, visual perception and biological vision and how these ideas relate to data augmentation and the contributions of this dissertation.

1.3.1 Supervised biological learning

The comparison between artificial intelligence—specifically artificial neural networks—and biological learning systems is intrinsic to the field, since one long-term goal of artificial intelligence is to mirror the capabilities of human intelligence. However, these capabilities are sometimes overestimated. One example is the argument that intelligence in nature evolves without supervision. In what follows, we will discuss three aspects of biological learning to argue against this view, in order to gain insights that better inform our progress in machine learning: first, we will discuss how generalisation requires exposure to relevant *training data*; second, the role of evolution and brain development; third, the variety of supervised signals that the brain may use.

First, in the argument that machine learning models should generalise from a few examples, there seems to be a promise or aspiration that with future better methods it will be possible for a neural network—for instance—to perform robust visual object categorisation among many object classes after being trained on one or a few examples per class. While we agree that a challenge for the near future is to *more efficiently* learn from fewer examples, we should also remind ourselves that no machine learning algorithm can robustly learn anything that cannot be inferred from the data it has been trained on. While this may seem to be against the ultimate goal of deep learning and a reason to look for radically different approaches, we should also bear in mind that learning in nature is not too different.

Even though the human visual system is remarkably robust, its capabilities are optimised for the tasks it needs to perform and largely shaped by experience, that is the *training data*—and years of evolution (Hasson et al., 2020), as we will discuss below. For instance, from the literature on human visual perception, we know that object recognition is viewpoint dependent (Tarr et al., 1998). A well-studied property of human vision is that our face recognition ability is severely impaired

if faces are presented upside down (Yin, 1969; Valentine, 1988). Setting aside the specific complexity of face processing in the brain, a compelling explanation for this impairment is that we are simply not used to seeing and recognising inverted faces. More generally, human perception of objects and our recognition ability is greatly affected when we see objects from unfamiliar viewpoints (Edelman & Bülthoff, 1992; Tarr et al., 1998; Bülthoff & Newell, 2006; Milivojevic, 2012). Furthermore, although better than the *one-shot* or *few-shot* generalisation of current ANNs, humans also have limited ability to recognise novel classes (Morgenstern et al., 2019). Interestingly, experiments with some novel classes of objects, known as Greebles, showed that with sufficient experience and training, humans can acquire expertise in recognising new objects from different viewpoints, even making use of an area of the brain—the fusiform face area—that typically responds strongly with face stimuli (Gauthier et al., 1999). This provides evidence that *generalisation* to multiple viewpoints is only developed after exposure to similar conditions. In this regard, data augmentation seems like a straightforward way to provide certain degree of input variability.

Second, the commonplace comparison of artificial neural networks with the brain often misses a fundamental component of biology, recently brought to the fore by Zador (2019) and Hasson et al. (2020), although considered since the early days of artificial intelligence (Turing, 1968): the role that millions of years of evolution have played in developing the nervous systems of organisms in nature, including the human brain. For example, some similarities have been observed between the representational geometry of the internal features learnt by ANNs trained for visual object recognition and that of the neural activations measured in the visual cortex of the brain (Yamins et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Güçlü & van Gerven, 2015). These articles were followed up by numerous studies that postulate ANNs as models of the visual cortex. However, what exactly drives better similarity with the brain remains an open question and this line of research presents several challenges.

While artificial neural networks are typically trained from scratch, from random initialisation, the neural activations are often measured in the adult brain. A reasonable approach would be to at least consider insights from developmental neuroscience and the infant brain (Harwerth et al., 1986; Atkinson, 2002; Gelman & Meyer, 2011). Moreover, as mentioned above, an interesting avenue is to also take into account the role of evolution. A large part of the brain connectivity is encoded genetically, and some properties of the visual cortex are known to be innate, that is without prior exposure to visual stimuli (Zador, 2019). Since neural networks are expected to learn some of these properties through extensive training on image data sets from scratch[6], part of the artificial learning process may have more to do with evolution than with the visual learning capabilities of an adult brain. This may be another reason to temper the expectations that neural networks be able to learn from a few examples, without *hard-wiring* some of the innate properties of the brain (Lindsey et al., 2019; Malhotra et al., 2020) or, alternatively, simulating part of the evolutionary process.

Third, another commonly found argument has it that children, and other animals in general, learn robust object recognition without supervision. First of all, we should recall again the role of evolution, which can be interpreted as a pre-trained model, optimised through millions of years of data with natural selection as a supervised signal (Zador, 2019). Second, we will argue against the very claim that children learn in fully unsupervised fashion. Obviously, the kind of supervision that humans make use of is not identical to that of machine learning algorithms: we do not see a class label on top of every object we look at. However, we receive supervision from multiple sources. Even though not for every visual stimulus, children do frequently receive information about the object classes they see—parents would point at objects and name them, for instance. Furthermore, humans usually follow a guided hierarchical learning: children do not directly learn to tell apart breeds of dogs, but rather start with umbrella terms and then progressively learn down the class hierarchy (Bornstein & Arterberry, 2010). Hasson et al. (2020) mention other examples of supervision from *social cues*, that is from other humans, such as learning to recognise

[6]Some interesting and promising areas in machine learning research deviate from this standard approach. For example, transfer learning studies and domain adaptation study the potential of features learnt on one task to be reused in different, related tasks (Zhuang et al., 2019), and continual learning studies the ways in which machine learning models can indefinitely sustain the acquisition of new knowledge without detriment of the previously learnt tasks (Aljundi, 2019; Mundt et al., 2019). These approaches are inspired by biological learning or share interesting properties with it.

individual faces, produce grammatical sentences, read and write; as well as from embodiment and action, such as learning to balance the body while walking or grasping objects. In all these actions, we can identify a supervised signal that surely influences learning in the brain (Shapiro, 2012).

Moreover, besides this kind of explicit supervision, the brain surely makes use of more subtle, implicit supervised signals, such as temporal stability (Becker, 1999; Wyss et al., 2003): The light that enters the retina is not a random signal from a sequence of rapidly changing arbitrary photos, but a highly coherent and regular flow of slowly changing stimuli, especially at the higher, semantical level (Kording et al., 2004). At the very least, this is how we perceive it and if such a smooth perception is a consequence instead of a cause, then it should be, again, a by-product of a long process of evolution. In the light of these insights from evolutionary theory and the explicit and implicit supervision that drives biological learning, we argue that we should temper the claims and aspirations that artificial neural networks should learn without supervision and from very few examples. Instead, we may benefit from rethinking the concept of supervision, embrace it and try to incorporate the forms of supervision present in nature into machine learning algorithms.

1.3.2 Supervised machine learning

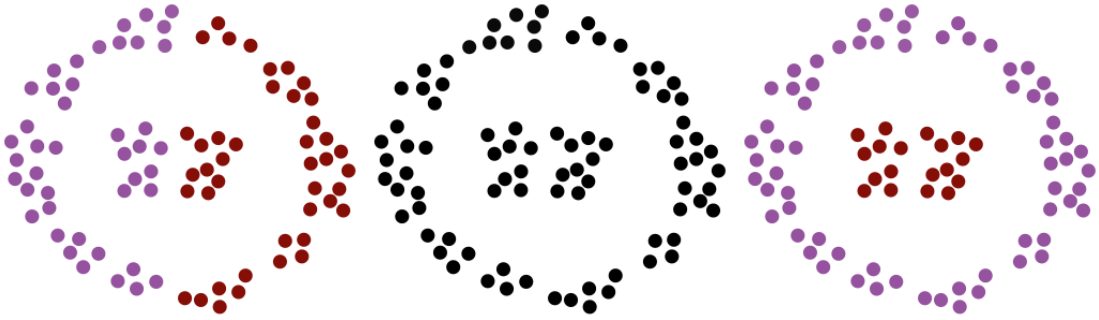
If we open a machine learning textbook (Murphy, 2012; Abu-Mostafa et al., 2012; Goodfellow et al., 2016), we will most surely find a taxonomy of learning algorithms with a clear distinction between *supervised* and *unsupervised* learning. If we take a look at the deep learning literature of the past years, we will also find abundant work on some variants *in between*: semi-supervised learning, self-supervised learning, etc. However, while this taxonomy can be useful, the boundaries are certainly not clear. As a matter of fact, strictly speaking, unsupervised learning is an illusion. If we recall the *no free lunch* theorem (Wolpert, 1996), averaged over all possible distributions, all classification algorithms are equivalent. Therefore, we need to constrain the distributions or, in other words, introduce prior knowledge—that is *supervision*. Recently, Locatello et al. (2018) obtained a similar result for the case of unsupervised learning of disentangled representations: without inductive biases for both the models and the data sets, unsupervised disentanglement learning is impossible. These results are purely theoretical and do not hold in real world applications, precisely because in practice we use multiple inductive biases, even when we do so-called unsupervised learning.

In a strict sense, even the classical, *purely* unsupervised methods, such as independent component analysis or nearest neighbours classifiers, make use of priors, such as independence or distance, respectively. Without inductive bias, learning is not possible. Consider the data points in Figure 1.2 (middle). With no prior information, all possible point categorisations are possible and equally valid. Depending on the inductive bias used, one model may find the configuration on the left, on the right, or any other. Which one is better depends on the task.

Although this is not news, the terminology used in the machine learning literature seems to neglect these nuances and evidences that the field suffers from *catastrophic forgetting*[7]. Particularly in deep learning and computer vision, the term *supervised* learning is often used to actually refer to *classification*, that is models trained on examples labelled according to the object classes, for instance. In turn, *unsupervised* learning is used for any model that does not use the labels, regardless of what other kind of supervision may be used. Further, the term *semi-supervised* learning refers to models that are trained with a fraction of the labels. While this terminology may be useful in some cases, it is not well defined and misses the fact that supervision can come in multiple flavours, not only as classification labels.

We have seen examples of different forms of supervision used by humans and other animals. What forms of supervision are common in machine learning? The relatively recent explosion of

[7] I am borrowing the expression from Prof. Irina Rish, who used it at a panel discussion of the UNIQUE Student Symposium 2020.



(a) One possible categorisation (b) A set of data points (c) One possible categorisation

Figure 1.2: An illustration of the need for inductive biases. Without any prior knowledge about the objective, any clustering of the data points is valid and can be potentially realised by a learning algorithm

deep learning has brought about the development of several libraries for automatic differentiation[8] (Baydin et al., 2017), which in turn have enabled the proposal of multiple loss functions with other types of supervision that can effectively be optimised by artificial neural networks through backpropagation and stochastic gradient descent. However, these approaches are often termed in the literature *semi-supervised*, *self-supervised* or even *unsupervised* learning. The term *semi-supervised* learning has gained popularity recently (Jing & Tian, 2020). On the one hand, this term acknowledges the fact that supervision is used—as opposed to unsupervised learning. On the other hand, it draws a hard line between classification and the other forms of supervised learning. The most likely reason for this separation is that labels are more costly to obtain than defining and implementing the tasks of self-supervised learning, rather than a formal, conceptual reason. From a theoretical point of view, both the conventional classification models and the recent wave of self-supervised objectives can be formalised within the category of supervised learning.

Importantly, the bulk of statistical learning theory (see a brief review in Section 2.1.2) has been developed for binary classification loss functions, then extended for multiclass classification, and in part for regression loss functions such as the mean squared error. Critically, the mapping of many results from statistical learning theory onto the various objective functions used in semi-supervised learning is far from trivial. Given the success of this kind of supervised objectives and their connection with perception, the study of these methods from a theoretical point of view might be a fruitful direction for future work.

1.4 Data augmentation, regularisation and inductive biases

After discussing the conflict in the terminology and acknowledging that purely unsupervised learning is an illusion, we can return to the role of inductive biases from visual perception and biological vision in defining useful forms of supervision for training artificial neural networks and, in particular, the role of data augmentation.

If we consider the particular, well-studied case of classification, to which the problem of visual object categorisation belongs, we can recall again the well-known *no free lunch* theorem, which establishes that learning is not possible without any prior. The field of statistical learning theory, whose fundamental results we review in Section 2.1.2, studies the conditions that make learning from data possible. One of its key results is that the space of possible solutions where an algorithm can search for a solution, the hypothesis set, has to be finite (Vapnik & Chervonenkis,

[8] Automatic differentiation is also known as algorithmic differentiation and differentiable programming, among other names. It is the set of techniques that allows to calculate the derivatives of numeric functions algorithmically. The development of automatic differentiation libraries has played a key role in the progress of deep learning in the past 10–15 years. Examples are Theano (Theano Development Team, 2016), TensorFlow (Abadi et al., 2015) and PyTorch (Paszke et al., 2019).

1971). Otherwise, the problem of inferring a function from a finite set of data is ill-posed. The classical way to ensure the problem is well-posed is through regularisation (Phillips, 1962; Tikhonov, 1963; Ivanov, 1976). Essentially, regularisation imposes a constraint on the hypothesis set—in the classical sense, a smoothness constraint—so that the inferred function cannot vary too rapidly around the training data points (see Section 2.1.3 for more details). Therefore, we can think of regularisation as an inductive bias.

Being such an essential ingredient of the learning problem, regularisation has been widely studied and multiple forms of regularisation have been proposed. Two ubiquitously used regularisation techniques in deep learning are weight decay and dropout (reviewed in Section 2.1.3). The former is a classical constraint on the norm of the learnable weights. The latter is a procedure to randomly turn off a subset of neurons during training. Both techniques have been shown to improve the generalisation of ANNs on the test data and hence their wide adoption. Thinking of regularisation techniques as inductive biases allows us to analyse the type of prior knowledge they incorporate, as well as widen the notion of regularisation to include other methods, such as data augmentation.

One of the contributions of this thesis is the comparison of weight decay, dropout and data augmentation. In Chapter 4, we show that neural networks trained on image object recognition tasks *without* weight decay and dropout achieve better performance than their counterparts, provided data augmentation is used during training. However, the common practice is to include all three: weight decay, dropout and data augmentation, among other methods. If we think of the inductive biases each technique introduces, data augmentation seems intuitively more advantageous: Weight decay assumes that models with smaller parameters should generalise better. Dropout assumes that neural networks that are forced to perform well with a subset of the neurons should perform better when all the neurons are used. Data augmentation makes use of an approximation of the oracle function, derived from prior knowledge about visual perception, that generates examples in regions of the input space where the model should learn a mapping. Although the inductive biases introduced by weight decay and dropout are clearly beneficial, the contribution of data augmentation seems more powerful. Nonetheless, these results were received with scepticism by the machine learning community.

In order to shed more light on the debate, we also draw theoretical insights from statistical learning theory that support the empirical findings and provide a grounded distinction between explicit regularisation methods, to which weight decay and dropout belong, and the implicit regularisation effect provided by data augmentation and other methods (Chapters 3 and 4). Explicit regularisation methods operate by directly constraining the hypothesis set of the model, known as representational capacity. However, many other methods that cannot be considered explicit regularisation provide an implicit regularisation effect, since they improve generalisation. Drawing a connection with the previously discussed ideas about supervision and inductive biases, we conclude that while explicit regularisation may indeed improve generalisation, it generally involves the optimisation of sensitive hyperparameters and at least the same or even better returns can be obtained by exploring ways of incorporating more meaningful inductive biases from visual perception and biological vision.

1.5 Invariance

In search of better inductive biases, we subscribe to the increasing trend in the machine community of exploring ways of training artificial neural networks beyond classification[9]. A pillar of this thesis is the attempt to integrate different disciplines. Therefore, in order to look for ways of improving visual object recognition models, we searched for inspiration in the mechanisms the brain has evolved in the visual cortex to solve object recognition.

The visual cortex is the part of the brain that processes visual information. One of its fundamental properties is the hierarchical organisation: while the primary areas of the visual cortex,

[9]Note we use the term *classification*, not *supervised learning*, after the discussion in Section 1.3.

which first process the information from the retina, respond to low-level properties such as the location in the visual field and the orientation of small parts of the stimuli (Hubel & Wiesel, 1962), the inferior temporal (IT) cortex, later in the processing pipeline, responds to higher-level properties such as the object category of the stimulus (Gross et al., 1972). This organisation of the biological networks in the brain greatly inspired the development of artificial neural networks.

Related to the hierarchical organisation of the visual cortex, Desimone et al. (1984) found that some neurons in the inferior temporal cortex of the macaque monkey responded consistently to the presentation of the same faces, regardless of their size and position. In contrast, neurons in earlier areas of the visual cortex are very sensitive to small changes in low-level properties of the visual stimuli, such as the orientation of edges. This invariance property was found to generalise to other objects besides faces (Booth & Rolls, 1998) in the late 1990s and in the 2000s was observed in the human brain too (Quiroga et al., 2005). Importantly, the invariance to identity-preserving transformations has been proposed to be an essential ingredient of robust visual object recognition in the brain (DiCarlo & Cox, 2007; Tacchetti et al., 2018). Hence, a reasonable question is whether artificial neural networks trained for visual object recognition are also invariant to such transformations.

The question of invariance in artificial neural networks has been addressed almost since their inception and studied from multiple perspectives. A large body of work has aimed at encoding different types of transformations into the networks (see a short review by Cohen & Welling (2016)), such as translation or rotation invariance. One contribution of this thesis is the study of the invariance of ANNs towards identity-preserving transformations using data augmentation. The use of data augmentation is convenient: Not coincidentally, the kind of stimulus transformations tested by computational neuroscientists to study the invariance to identity-preserving transformations in the inferior temporal cortex and the image transformations typically included in data augmentation schemes are the same, or very similar. These are the transformations that we encounter naturally in the visual world, as we perceive it along the temporal dimension (Kording et al., 2004; Einhäuser et al., 2005; Taylor et al., 2011), which perceptually preserve the object identity. For this reason we have named them *perceptually plausible* transformations.

First, we measured the invariance to identity-preserving transformations of models trained on image object recognition data sets (Chapter 6). Intuitively and also taking the insights from the properties of the IT cortex, two images that represent different views of the same object should produce similar activations at the higher layers of a neural network. However, we found that the similarity is not better than at the pixel space, even though the models do classify the images correctly and were exposed to such transformations during training. This finding contradicts in part the general intuition that neural networks learn hierarchical representations, ranging from specific to more abstract, object-related features. We hypothesised that this is a sign of a lack of perceptual inductive bias. Given the large representational capacity of neural networks, training them with the sole objective of classifying a data set of images into the right classes does not seem enough to learn perceptually invariant representations, despite the theoretical results showing that invariance to *nuisances* should emerge naturally (Achille & Soatto, 2018). In general, there exist multiple possible solutions for the classification problem within the hypothesis space spanned by modern deep neural networks and the models do not seem to naturally converge to solutions well aligned with some crucial aspects of visual perception and biological vision (Sinz et al., 2019; Geirhos et al., 2020; Dujmović et al., 2020). This led us to propose *data augmentation invariance*, an objective function that encourages robust representations, inspired by the invariance observed in the visual cortex.

In Chapter 6, we discuss the details of data augmentation invariance. We trained artificial neural networks on image object recognition data sets by jointly optimising the categorisation objective and a new data augmentation invariance objective. The latter is a layer-wise objective that encourages that the representations of transformations of the same image—generated through perceptually plausible data augmentation—cluster together. We attempted to simulate the increasing invariance along the visual cortex hierarchy by distributing the invariance loss exponentially along the neural network layers. Models trained with data augmentation invariance effectively learnt increasingly invariant representations without detriment to the classification ac-

curacy, which even improved in some cases.

In view of these results we argue that replacing or complementing the standard classification objectives with perceptually and biologically inspired objectives, such as data augmentation invariance, is a promising avenue to both improve computer vision algorithms and obtain better models of natural vision. Such objectives are likely not biologically plausible, in the sense that the brain does not optimise an equivalent objective (Pozzi et al., 2018). However, since properties like invariance to identity-preserving transformations are at least a by-product of either evolution or early brain development (or both), it is reasonable and potentially fruitful to optimise artificial neural networks with objectives that simulate key properties of the brain.

1.6 Visual salience

So far our discussion around visual perception, biological vision and machine image understanding algorithms has revolved mainly around object recognition. However, animal vision encompasses a broader range of capabilities that allow us to navigate and understand the world. As part of the interdisciplinary pursuit of this work, we here studied some aspects of another central component of vision: visual attention.

Visual attention is a complex brain mechanism that enables us to coherently process the sheer amount of light that enters our eyes. At any given time, even though the retina receives stimulation from the whole visual field, only a small fraction of the information is processed in detail (Desimone & Duncan, 1995). Specifically, the visual system preferentially processes the information located in the centre of the visual field—the *fovea*—as the central part of the retina has a larger density of photoreceptors than in the surroundings—the *visual periphery* (Wässle et al., 1990; Azzopardi & Cowey, 1993). For instance, a recent study has shown that many people fail to notice when up to 95 % of the visual field is presented without colour (Cohen et al., 2020). What particular area of the available information in front of us is processed in most detail at a time is mediated by eye movements, and what exactly drives eyes movement is a complex, widely studied question, which remains largely open.

For example, it is known that eye movements can be driven by both low-level properties of the stimuli—*bottom-up*—and by cognitive processes derived from, for instance, a task or desire—*top-down* (Von Stein et al., 2000; Munoz & Everling, 2004; Connor et al., 2004; Betz et al., 2010; Kollmorgen et al., 2010; Schütt et al., 2019). An interesting subject of study is the relationship between object recognition and visual attention. There is strong evidence for the role of object recognition in the direction of eye movements (Zhaoping & Guyader, 2007), but visual attention has been also suggested to predict object perceptual awareness (Holm et al., 2008; Kietzmann et al., 2011).

From a behavioural perspective, the majority of the research work that studies visual attention makes use of eye tracking devices, which are able to map the gaze of an observer at any given time with the location on a stimulus. Another active area of research, at the intersection between vision science and computer vision, is the modelling of visual salience, first proposed by Itti et al. (1998). Salience models shift the focus to the stimulus side and aim to answer the question "what parts of a stimulus are most salient to a human observer?". Adhering to the definitions by Kümmerer et al. (2018), a salience model predicts the probability that a pixel on a given image is fixated, which can be expressed through salience maps that represent the distribution of salience for specific tasks. For this thesis we made use of both eye tracking and salience maps to study some aspects of human vision.

In one project, presented in detail in Chapter 7, we were interested in studying the global salience of competing stimuli, that is stimuli presented side by side. The bulk of the research on computational models of visual salience addresses the question of what parts of a stimulus are more likely to attract the gaze of observers. In this case, we aimed at quantifying the salience of images as a whole, to seek answers for the questions: Are some types of images more likely to attract the gaze of observers? If so, is this global salience related to the local salience properties of

the images? Do other factors, such as familiarity with one of the images, play a role in the gaze direction of observers? In order to answer these questions, we conducted eye tracking experiments in which we recorded the gaze direction of participants who were shown pairs of images side by side. We then modelled the behavioural data with a machine learning algorithm and computed the local salience properties of the images with representative salience models from the literature. As a main finding, we concluded that natural images intrinsically have a global salience that varies widely across different types of images and is independent of the local salience properties.

In another study, we combined computational models of visual salience with brain measurements of functional magnetic resonance imaging (fMRI) to analyse properties of the human visual cortex. We followed up the work by Zuiderbaan et al. (2017), where the authors showed that it is possible to identify which natural image was shown to a participant in the scanner from fMRI recordings. The predictions were made by comparing the brain activations elicited by each image on areas V1, V2 and V3, and a combination of a contrast map of the images with the receptive field properties of the cortical areas, obtained through the population receptive field (pRF) model (Dumoulin & Wandell, 2008). Here, we studied whether brain activity was better predicted by salience maps than by contrast maps, and extended the analysis to a broader range of visual cortical areas: V1, V2, V3, hV4, LO12 and V3AB (Wandell et al., 2007). We studied two distinct models of visual salience, ICF and DeepGaze (Kümmerer et al., 2017), and concluded that salience is more predictive of brain activations than contrast, especially the salience model based on intensity and contrast information only (ICF), rather than on high-level features, suggesting that salience information is still present in the neural activations of the visual cortex.

1.7 Overview of contributions and outline

The overarching objective of this thesis is to explore and exploit the connections between deep artificial neural networks and the visual cognitive and neural sciences. We believe that all three fields can benefit from mutual collaboration and synergies.

In order to facilitate the understanding of the thesis to a broader audience and set the grounds for the discussion throughout the dissertation, Chapter 2 provides an introduction to the fundamentals of both machine learning and visual object recognition in the brain, as well as other relevant concepts. This chapter serves also as a review of related scientific literature.

Then, a first block from Chapter 3 to 6, has data augmentation as the central theme, starting from the rather machine learning-centred Chapter 3, towards gradually incorporating aspects from visual perception and biological vision in the subsequent chapters.

Specifically, in Chapter 3, we discuss the concepts of explicit and implicit regularisation and provide definitions of these terms that have been widely but ambiguously used in the literature. Part of this chapter is based on the publication (Hernández-García & König, 2018b) and we here provide a longer discussion about the taxonomy of regularisation and examples of explicit and implicit regularisation, arguing in particular that data augmentation is not explicit regularisation, as considered before in the literature.

Chapter 4 is focused on the comparison of explicit regularisation techniques—weight decay and dropout—and data augmentation and much of the content has been published in several articles (Hernández-García & König, 2018a;b;c). We present the results of a systematic empirical evaluation, alongside some insights from statistical learning theory, to conclude that data augmentation alone can provide the same generalisation gain than combined with explicit regularisation, and is remarkably more flexible. In view of these results, we discuss the need for weight decay and dropout in deep learning and propose to rethink the status of data augmentation.

In Chapter 5, we compare the representations learnt by neural networks trained with data augmentation and the activations in the inferior temporal cortex of the human brain (Hernández-García et al., 2018). We found that models trained with heavier transformations learn features more aligned with the representations in the higher visual cortex.

Following up the connection between data augmentation and biological vision, in Chapter 6, we study one of the fundamental properties of the visual cortex: the increasing invariance along the ventral stream to identity-preserving transformations of visual objects. Using data augmentation as a framework to generate such transformations, we first show that standard artificial neural networks trained optimised for object categorisation are hardly robust in terms of representational similarity. Then, we propose *data augmentation invariance* as a simple, yet effective and efficient way of learning robust features, while preserving the categorisation performance (Hernández-García et al., 2019b).

The last two chapters of the dissertation can be seen as a separate block, in which data augmentation and artificial neural networks are not the main subject, although machine learning is still used as a tool. Chapter 7 is closer to the field of cognitive science, as we study some aspects of visual behaviour through an eye-tracking experiment (Hernández-García et al., 2019a). In particular, we propose *global visual salience* as a metric of the likelihood of competing natural images to attract the gaze of an observer. Chapter 8 is closer to neuroimaging, as we compare models of image identification from brain data to study properties of the early visual cortex. Part of the results of this study were presented as a poster contribution at the Annual Meeting of the Visual Sciences Society in 2019 (Hernández-García et al., 2019c), and we here extended the analysis.

We conclude the dissertation with a general discussion in Chapter 9, where we provide an overview of the main results, outline the connections between the different parts, discuss future lines of work and the broader potential impact of this work.

Bibliography

Abadi, M. et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

Abu-Mostafa, Y. S. Learning from hints in neural networks. *Journal of Complexity*, 1990.

Abu-Mostafa, Y. S., Magdon-Ismail, M., and Lin, H.-T. *Learning from data*. AMLBooks, 2012.

Achille, A. and Soatto, S. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research (JMLR)*, 2018.

Aljundi, R. *Continual Learning in Neural Networks*. PhD thesis, KU Leuven, Faculty of Engineering Science, 2019.

Atkinson, J. *The developing visual brain*. Oxford Scholarship Online, 2002.

Audebert, N., Le Saux, B., and Lefèvre, S. Deep learning for classification of hyperspectral data: A comparative review. *IEEE Geoscience and Remote Sensing Magazine*, 2019.

Azzopardi, P. and Cowey, A. Preferential representation of the fovea in the primary visual cortex. *Nature*, 1993.

Baydin, A. G., Pearlmutter, B. A., Radul, A. A., and Siskind, J. M. Automatic differentiation in machine learning: a survey. *Journal of Machine Learning Research (JMLR)*, 2017.

Becker, S. Implicit learning in 3D object recognition: The importance of temporal context. *Neural Computation*, 1999.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences (PNAS)*, 2019.

Bengio, Y., Lee, D.-H., Bornschein, J., Mesnard, T., and Lin, Z. Towards biologically plausible deep learning. *arXiv preprint arXiv:1502.04156*, 2015.

Betz, T., Kietzmann, T. C., Wilming, N., and König, P. Investigating task-dependent top-down effects on overt visual attention. *Journal of Vision*, 2010.

Booth, M. and Rolls, E. T. View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cerebral Cortex*, 1998.

Bornstein, M. H. and Arterberry, M. E. The development of object categorization in young children: Hierarchical inclusiveness, age, perceptual attribute, and group versus individual analyses. *Developmental Psychology*, 2010.

Bowers, J. S. Parallel distributed processing theory in the age of deep networks. *Trends in Cognitive Sciences*, 2017.

Bülthoff, I. and Newell, F. N. The role of familiarity in the recognition of static and dynamic objects. *Progress in Brain Research*, 2006.

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. VGGFace2: A dataset for recognising faces across pose and age. In *IEEE International Conference on Automatic Face & Gesture Recognition*. 2018.

Cohen, M. A., Botch, T. L., and Robertson, C. E. The limits of color awareness during active, real-world vision. *Proceedings of the National Academy of Sciences (PNAS)*, 2020.

Cohen, T. and Welling, M. Group equivariant convolutional networks. In *International Conference on Machine Learning (ICML)*, 2016.

Connor, C. E., Egeth, H. E., and Yantis, S. Visual attention: bottom-up versus top-down. *Current Biology*, 2004.

Cortes, C. and Vapnik, V. Support-vector networks. *Machine Learning*, 1995.

Dalal, N. and Triggs, B. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2005.

Desimone, R. and Duncan, J. Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 1995.

Desimone, R., Albright, T. D., Gross, C. G., and Bruce, C. Stimulus-selective properties of inferior temporal neurons in the macaque. *Journal of Neuroscience*, 1984.

DeVries, T. and Taylor, G. W. Dataset augmentation in feature space. In *International Conference on Learning Representations (ICLR), arXiv:1702.05538*, 2017.

DiCarlo, J. J. and Cox, D. D. Untangling invariant object recognition. *Trends in Cognitive Sciences*, 2007.

Douglas, R. J., Martin, K. A., and Whitteridge, D. A canonical microcircuit for neocortex. *Neural Computation*, 1989.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research (JMLR)*, 2011.

Duda, R. O. and Hart, P. E. Use of the Hough transformation to detect lines and curves in pictures. *Communications of the ACM*, 1972.

Dujmović, M., Malhotra, G., and Bowers, J. What do adversarial images tell us about human vision? *bioRxiv preprint 2020.02.25.964361*, 2020.

Dumoulin, S. O. and Wandell, B. A. Population receptive field estimates in human visual cortex. *Neuroimage*, 2008.

Edelman, S. and Bülthoff, H. H. Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. *Vision Research*, 1992.

Einhäuser, W., Hipp, J., Eggert, J., Körner, E., and König, P. Learning viewpoint invariant object representations using a temporal coherence principle. *Biological Cybernetics*, 2005.

Fukushima, K. and Miyake, S. Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. *Pattern Recognition*, 1982.

Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., and Gore, J. C. Activation of the middle fusiform'face area'increases with expertise in recognizing novel objects. *Nature Neuroscience*, 1999.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *arXiv preprint arXiv:2004.07780*, 2020.

Gelman, S. A. and Meyer, M. Child categorization. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2011.

Gencoglu, O., van Gils, M., Guldogan, E., Morikawa, C., Süzen, M., Gruber, M., Leinonen, J., and Huttunen, H. HARK side of deep learning–from grad student descent to automated machine learning. *arXiv preprint arXiv:1904.07633*, 2019.

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

Glorot, X., Bordes, A., and Bengio, Y. Deep sparse rectifier neural networks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

Gonzalez, R. and Woods, R. *Digital image processing*. Pearson, 4th edition, 2018.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep learning*. MIT press, 2016.

Goodfellow, I. J., Warde-Farley, D., Mirza, M., Courville, A. C., and Bengio, Y. Maxout networks. In *International Conference on Machine Learning (ICML)*, 2013.

Graham, B. Fractional max-pooling. *arXiv preprint arXiv:1412.6071*, 2014.

Gross, C. G., Rocha-Miranda, C. d., and Bender, D. Visual properties of neurons in inferotemporal cortex of the macaque. *Journal of Neurophysiology*, 1972.

Güçlü, U. and van Gerven, M. A. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 2015.

Harris, K. D. and Shepherd, G. M. The neocortical circuit: themes and variations. *Nature Neuroscience*, 2015.

Harwerth, R. S., Smith, E. L., Duncan, G. C., Crawford, M., and Von Noorden, G. K. Multiple sensitive periods in the development of the primate visual system. *Science*, 1986.

Hassabis, D., Kumaran, D., Summerfield, C., and Botvinick, M. Neuroscience-inspired artificial intelligence. *Neuron*, 2017.

Hasson, U., Nastase, S. A., and Goldstein, A. Direct fit to nature: An evolutionary perspective on biological and artificial neural networks. *Neuron*, 2020.

Haugeland, J. *Artificial intelligence: The very idea*. MIT Press, 1989.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Hebb, D. O. *The organization of behavior: a neuropsychological theory*. Wiley, 1949.

Hernández-García, A. Aesthetics assessment of videos through visual descriptors and automatic polarity annotation. B.S. thesis, Universidad Carlos III de Madrid, 2014.

Hernández-García, A. and König, P. Further advantages of data augmentation on convolutional neural networks. In *International Conference on Artificial Neural Networks (ICANN)*. 2018a.

Hernández-García, A. and König, P. Data augmentation instead of explicit regularization. *arXiv preprint arXiv:1806.03852*, 2018b.

Hernández-García, A. and König, P. Do deep nets really need weight decay and dropout? *arXiv preprint arXiv:1802.07042*, 2018c.

Hernández-García, A., Fernández-Martínez, F., and Díaz-de María, F. Emotion and attention: Predicting electrodermal activity through video visual descriptors. In *International Conference on Web Intelligence*, 2017.

Hernández-García, A., Mehrer, J., Kriegeskorte, N., König, P., and Kietzmann, T. C. Deep neural networks trained with heavier data augmentation learn features closer to representations in hIT. In *Conference on Cognitive Computational Neuroscience (CCN)*, 2018.

Hernández-García, A., Gameiro-Ramos, R., Grillini, A., and König, P. Global visual salience of competing stimuli. *PsyArXiv preprint PsyArXiv:z7qp5*, 2019a.

Hernández-García, A., König, P., and Kietzmann, T. Learning robust visual representations using data augmentation invariance. *arXiv preprint arXiv:1906.04547*, 2019b.

Hernández-García, A., Zuiderbaan, W., Edadan, A., Dumoulin, S. O., and König, P. Saliency and the population receptive field model to identify images from brain activity. In *Annual Meeting of the Vision Sciences Society (VSS)*. 2019c.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural Computation*, 1997.

Holm, L., Eriksson, J., and Andersson, L. Looking as if you know: Systematic object inspection precedes object recognition. *Journal of Vision*, 2008.

Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 1991.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

Hubel, D. H. and Wiesel, T. N. Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, 1959.

Hubel, D. H. and Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 1962.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, 2015.

Itti, L., Koch, C., and Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 1998.

Ivanov, V. V. *The theory of approximate methods and their applications to the numerical solution of singular integral equations*. Nordhof International, 1976.

Jing, L. and Tian, Y. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

Khaligh-Razavi, S.-M. and Kriegeskorte, N. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLOS Computational Biology*, 2014.

Kietzmann, T. C., Geuter, S., and König, P. Overt visual attention as a causal factor of perceptual awareness. *PLOS ONE*, 2011.

Kietzmann, T. C., McClure, P., and Kriegeskorte, N. Deep neural networks in computational neuroscience. *Oxford Research Encyclopedia of Neuroscience*, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kollmorgen, S., Nortmann, N., Schröder, S., and König, P. Influence of low-level stimulus features, task dependent factors, and spatial biases on overt visual attention. *PLOS Computational Biology*, may 2010.

Kording, K. P., Kayser, C., Einhauser, W., and Konig, P. How are complex cell properties adapted to the statistics of natural stimuli? *Journal of Neurophysiology*, 2004.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. *Technical report, University of Toronto*, 2009.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012.

Kümmerer, M., Wallis, T. S., Gatys, L. A., and Bethge, M. Understanding low-and high-level contributions to fixation prediction. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

Kümmerer, M., Wallis, T. S., and Bethge, M. Saliency benchmarking made easy: Separating models, maps and metrics. In *European Conference on Computer Vision (ECCV)*, 2018.

Larsson, G., Maire, M., and Shakhnarovich, G. FractalNet: Ultra-deep neural networks without residuals. *arXiv preprint arXiv:1605.07648*, 2016.

Latif, A., Rasheed, A., Sajid, U., Ahmed, J., Ali, N., Ratyal, N. I., Zafar, B., Dar, S. H., Sajid, M., and Khalil, T. Content-based image retrieval and feature extraction: a comprehensive review. *Mathematical Problems in Engineering*, 2019.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature*, 2015.

Lienhart, R. and Maydt, J. An extended set of Haar-like features for rapid object detection. In *International Conference on Image Processing (ICIP)*. 2002.

Lindsay, G. Convolutional neural networks as a model of the visual system: past, present, and future. *Journal of Cognitive Neuroscience*, 2020.

Lindsay, G. W. and Miller, K. D. How biological attention mechanisms improve task performance in a large-scale visual system model. *eLife*, 2018.

Lindsey, J., Ocko, S. A., Ganguli, S., and Deny, S. A unified theory of early visual representations from retina to cortex through anatomically constrained deep CNNs. In *International Conference on Learning Representations (ICLR), arXiv:1901.00945*, 2019.

Locatello, F., Bauer, S., Lucic, M., Rätsch, G., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. *arXiv preprint arXiv:1811.12359*, 2018.

Logothetis, N. K. and Sheinberg, D. L. Visual object recognition. *Annual Review of Neuroscience*, 1996.

Lowe, D. G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 2004.

Malhotra, G., Evans, B. D., and Bowers, J. S. Hiding a plane with a pixel: examining shape-bias in CNNs and the benefit of building in biological constraints. *Vision Research*, 2020.

Marblestone, A. H., Wayne, G., and Kording, K. P. Toward an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience*, 2016.

Marcus, G. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018.

McCulloch, W. S. and Pitts, W. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 1943.

Milivojevic, B. Object recognition can be viewpoint dependent or invariant–it's just a matter of time and task. *Frontiers in Computational Neuroscience*, 2012.

Morgenstern, Y., Schmidt, F., and Fleming, R. W. One-shot categorization of novel object classes in humans. *Vision Research*, 2019.

Mundt, M., Majumder, S., Pliushch, I., and Ramesh, V. Unified probabilistic deep continual learning through generative replay and open set recognition. *arXiv preprint arXiv:1905.12019*, 2019.

Munoz, D. P. and Everling, S. Look away: the anti-saccade task and the voluntary control of eye movement. *Nature Reviews Neuroscience*, 2004.

Murphy, K. P. *Machine learning: a probabilistic perspective*. MIT Press, 2012.

Nayebi, A. and Ganguli, S. Biologically inspired protection of deep networks from adversarial attacks. *arXiv preprint arXiv:1703.09202*, 2017.

Paszke, A. et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2019.

Perronnin, F. and Dance, C. Fisher kernels on visual vocabularies for image categorization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2007.

Phillips, D. L. A technique for the numerical solution of certain integral equations of the first kind. *Association for Computing Machinery*, 1962.

Pozzi, I., Bohté, S., and Roelfsema, P. A biologically plausible learning rule for deep learning in the brain. *arXiv preprint arXiv:1811.01768*, 2018.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., and Fried, I. Invariant visual representation by single neurons in the human brain. *Nature*, 2005.

Richards, B. A. et al. A deep learning framework for neuroscience. *Nature Neuroscience*, 2019.

Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 1958.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning representations by back-propagating errors. *Nature*, 1986.

Russakovsky, O. et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 2015.

Saxe, A., Nelli, S., and Summerfield, C. If deep learning is the answer, then what is the question? *arXiv preprint arXiv:2004.07580*, 2020.

Schütt, H. H., Rothkegel, L. O., Trukenbrod, H. A., Engbert, R., and Wichmann, F. A. Disentangling bottom-up versus top-down and low-level versus high-level influences on eye movements over time. *Journal of Vision*, 2019.

Shapiro, L. *Embodied cognition*. Oxford Handbooks Online, 2012.

Simard, P., Victorri, B., LeCun, Y., and Denker, J. Tangent prop-a formalism for specifying selected invariances in an adaptive network. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1992.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Sinz, F. H., Pitkow, X., Reimer, J., Bethge, M., and Tolias, A. S. Engineering a less artificial intelligence. *Neuron*, 2019.

Sivic, J. and Zisserman, A. Video google: A text retrieval approach to object matching in videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2003.

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. Striving for simplicity: The all convolutional net. In *International Conference on Learning Representations (ICLR), arXiv:1412.6806*, 2014.

Tacchetti, A., Isik, L., and Poggio, T. A. Invariant recognition shapes neural representations of visual input. *Annual Review of Vision Science*, 2018.

Tarr, M. J., Williams, P., Hayward, W. G., and Gauthier, I. Three-dimensional object recognition is viewpoint dependent. *Nature Neuroscience*, 1998.

Taylor, G. W., Spiro, I., Bregler, C., and Fergus, R. Learning invariance through imitation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2011.

Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv preprint arXiv:1605.02688*, 2016.

Tikhonov, A. N. On solving ill-posed problem and method of regularization. *Doklady Akademii Nauk USSR*, 1963.

Turing, A. M. *Cybernetics; (Key papers)*. University Park Press, 1968.

Valentine, T. Upside-down faces: A review of the effect of inversion upon face recognition. *British Journal of Psychology*, 1988.

Van Essen, D. C. Organization of visual areas in macaque and human cerebral cortex. *The Visual Neurosciences*, 2003.

Vapnik, V. N. and Chervonenkis, A. Y. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 1971.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Von Stein, A., Chiang, C., and König, P. Top-down processing mediated by interareal synchronization. *Proceedings of the National Academy of Sciences (PNAS)*, 2000.

Wandell, B. A., Dumoulin, S. O., and Brewer, A. A. Visual field maps in human cortex. *Neuron*, 2007.

Wässle, H., Grünert, U., Röhrenbeck, J., and Boycott, B. B. Retinal ganglion cell density and cortical magnification factor in the primate. *Vision Research*, 1990.

Wolpert, D. H. The lack of a priori distinctions between learning algorithms. *Neural Computation*, 1996.

Wyss, R., König, P., and Verschure, P. F. Invariant representations of visual patterns in a temporal population code. *Proceedings of the National Academy of Sciences (PNAS)*, 2003.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences (PNAS)*, 2014.

Yin, R. K. Looking at upside-down faces. *Journal of Experimental Psychology*, 1969.

Zador, A. M. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature Communications*, 2019.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR), arXiv:1611.03530*, 2017a.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017b.

Zhao, Z.-Q., Zheng, P., Xu, S.-t., and Wu, X. Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems*, 2019.

Zhaoping, L. and Guyader, N. Interference with bottom-up feature detection by higher-level object recognition. *Current Biology*, 2007.

Zhou, W., Li, H., and Tian, Q. Recent advance in content-based image retrieval: A literature survey. *arXiv preprint arXiv:1706.06064*, 2017.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. A comprehensive survey on transfer learning. *arXiv preprint arXiv:1911.02685*, 2019.

Zoph, B. and Le, Q. V. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016.

Zuiderbaan, W., Harvey, B. M., and Dumoulin, S. O. Image identification from brain activity using the population receptive field model. *PLOS ONE*, 2017.

Chapter 2

Background

The purpose of this chapter is two-fold: First, it aims at providing the fundamental background about the most relevant aspects of machine learning for this thesis, such as the theory of generalisation and regularisation. Second, it is also intended to serve as a review of the relevant and related scientific literature.

The introduction to the core aspects of machine learning in this chapter is deliberately non-exhaustive, as it is intended to provide only a sufficient background to enable the understanding of the rest of the thesis to a wider audience. The interested reader may follow the references to scientific literature provided throughout the chapter.

2.1 Machine learning fundamentals

Humans perceive the world and make sense of it in a great variety of ways: we see light as visual information, hear sounds and create music, use language to produce written text and speech and organise much of what we know into collections of numbers, to name a few. As diverse as light, sound, language and numbers are, they can all be conceptualised as *data*. Thinking of what we perceive and know about the world as information that can be organised into data is a powerful formal conceptualisation that allows us to better understand and analyse the input and output we perceive and generate.

Machine learning is the discipline that studies how to automatically discover patterns from data (Murphy, 2012). Much as humans and other animals are able to make sense of the world out of what they perceive, machine learning aims at providing the methods to make sense out of formally defined data points, that is learning from data (Abu-Mostafa et al., 2012). Machine learning has its roots in mathematics and statistical inference, but has developed as a distinct field after the development and spread of computers and computer science, which have provided the means to store and process data efficiently and automatically.

2.1.1 Elements of machine learning

The fundamental component of machine learning is the data and, as motivated above, the field itself arises from the ability to map any observation of the world into data points. Formally, one can define one data point as an M-dimensional vector $\mathbf{x} = x_j, \ldots, x_M$. Then, a set of N observations $\{\mathbf{x}_i\}_i^N \in \mathcal{X}$ is said to be the *input* data set. Most of the machine learning literature (Alpaydin, 2009; Abu-Mostafa et al., 2012; Murphy, 2012) makes a broad distinction amongst machine learning methods depending on the *output* or *target* data. A non-exhaustive taxonomy of the main types of machine learning methods as commonly found in the literature is the following:

- **Supervised learning**: in a supervised learning setting, every observation $\mathbf{x}_i$ is paired with an output variable y_i, also referred to as *ground truth*, and thus the data set is considered $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_i^N$. The goal of supervised learning is discovering the relationship between the input data $\mathbf{x}_i$ and the target variables y_i. Depending on the nature of y_i, the learning problem can be *classification* or *regression*:
  - Classification: $y_i \in \{1, \dots, C\}$ is a discrete or categorical value. The possible values of y are also referred to as *classes* or *labels*.
  - Regression: $y_i \in \mathbb{R}$ is a continuous variable.
- **Unsupervised learning**: in an unsupervised learning setting, there is not explicit access to target variables. Therefore the data set is simply $\mathcal{D} = \{\mathbf{x}_i\}_i^N$, and the goal is to discover patterns in the input data. A prototypical example of unsupervised learning is clustering.
- **Reinforcement learning** is a broad class of machine learning methods, initially inspired by behavioural psychology and the concept of trial-and-error learning. Instead of a mapping between input and output variables, in reinforcement learning typically there is access to a *reward* signal that might not be available for every input data point. The goal is to learn policy that maximises the expected rewards by seeking a balance between exploration—the acquisition of new knowledge—and exploitation—the use of that knowledge to improve performance.

While this distinction is useful, the boundaries are sometimes blurred, in practice. For example, some problems often labelled as unsupervised learning could be considered particular cases of supervised learning, as we have discuss in the Introduction (Section 1.3). Most of the problems that we will address in this thesis can be defined within a supervised framework and, more specifically, classification. For these reasons, in the remaining of this section we will focus on classification, unless specified otherwise.

As introduced above, the goal of a (supervised) machine learning method is to discover the relationship between the input data and the target variables. This assumes that such relationship is determined by an underlying, unknown function $f \colon \mathcal{X} \mapsto \mathcal{Y}$. Since f is a latent function, the task is to find a function $g \in \mathcal{H} \colon \mathcal{X} \mapsto \mathcal{Y}$ from a set of candidate functions $h \in \mathcal{H}$—the hypothesis set—that approximates f according to certain error or loss measure $L(h, f)$. In order to find g, the *learning algorithm* $\mathcal{A}$ uses the available *training* data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_i^N = (\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)$ to solve an optimisation problem by adjusting a set of learnable parameters $\boldsymbol{\theta}$. Hence, the task is to determine from the set of functions $h(\mathbf{x}; \boldsymbol{\theta})$, the one which best approximates the data $\mathcal{D}$.

Nonetheless, if the relationships found apply only to the training data $\mathcal{D}$, then the process could not be considered *learning*, but at best memorisation. Crucially, the ultimate objective of machine learning is to learn relationships and make correct predictions beyond the observed data. This is called *generalisation*. This notion of learning is only feasible in a probabilistic way. The probabilistic view introduces the important assumption that the relationship between the targets y and the input $\mathbf{x}$ is not deterministic, but probabilistic and there exists an unknown, underlying joint probability distribution $P_{X,Y}(\mathbf{x}, y)$ on $\mathcal{X} \times \mathcal{Y}$—thus also a marginal input distribution $P_X(\mathbf{x})$ and a conditional output distribution $P_{Y|X}(y|\mathbf{x})$, in Bayesian terms. Furthermore, it is also generally assumed that the observed available data points were sampled independently from $P_{X,Y}$. A summary schematic of the main elements of supervised learning is shown in Figure 2.1.

The feasibility of learning from a mathematical and probabilistic perspective is studied by the field of statistical learning theory (Vapnik, 1995; Bousquet et al., 2003; Von Luxburg & Schölkopf, 2011). In the next section, we introduce the important concept of generalisation in machine learning and some notions from learning theory that are relevant to this thesis.

2.1.2 Theory of generalisation

Generalisation is one of the most important concepts in machine learning and statistical learning theory. It refers to the idea that the ultimate goal in the learning process is not to minimise the

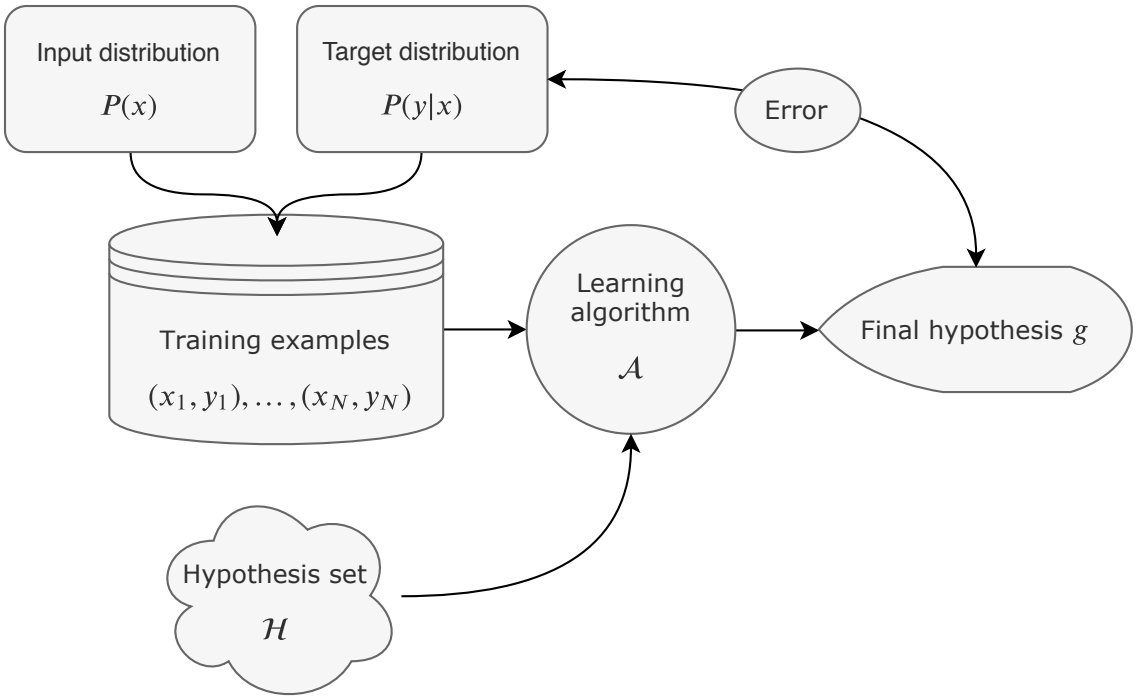


Figure 2.1: Schematic of the main elements of (supervised) learning. Adapted from Abu-Mostafa et al. (2012).

error computed on the available training data in and of itself, but to perform well on unseen data, that is to generalise. This raises the reasonable first question of whether generalisation is possible at all. The probabilistic perspective not only provides a positive answer to the question but also the tools to analyse the generalisation guarantees of learning algorithms.

Empirical risk minimisation

In order to elaborate this idea, let us first introduce some important concepts that we will use in this section and throughout the thesis. To formally describe the problem at hand, we recall that we consider data that consist of the inputs $\mathbf{x}_i \in \mathcal{X}$ and the outputs or targets $y_i \in \mathcal{Y} = \{-1, 1\}$, that is binary classification, for simplicity of the exposition. Furthermore, we assume that the pairs $(\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ are independently and identically sampled according to an unknown probability distribution $P_{X,Y}$. So as to measure the discrepancy between the target variables y and the outcome of the hypotheses $h(x)$,[1] we assume that we are given a real-valued *loss function* $L(y, h(x))$. Since we are considering binary classification, the loss function will be the classification error: $L(y, h(x)) = \mathbb{1}_{h(x) \neq y}$. Then, the *risk* associated with a hypothesis h is given by the expectation of the loss function, defined by the following *risk functional*:

$$R(h) = \mathbb{E}\left[L(y, h(x))\right] = \int L(y, h(x)) dP_{X,Y}(x, y) \tag{2.1}$$

Ultimately, the goal of a learning algorithm is to find the optimal hypothesis $h^* \in \mathcal{H}$ that minimises the risk $R(h)$:

$$h^* = \arg\min_{h \in \mathcal{H}} R(h) \tag{2.2}$$

However, because the joint probability distribution $P_{X,Y}$ is unknown, it is not possible to exactly calculate $R(h)$. In practice, the risk functional is replaced by the computation of the *empirical*

[1] The hypotheses depend on both $\mathbf{x}$ and $\boldsymbol{\theta}$, that is $h(\mathbf{x}; \boldsymbol{\theta})$, but we will in general abuse notation and write simply $h(x)$, or even just h, for better readability.

risk on the set of N available data points:

$$R_N(h) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, h(x_i, \theta)) \tag{2.3}$$

and the learning algorithm chooses the hypothesis g by minimising the empirical risk:

$$g = \hat{h} = \arg\min_{h \in \mathcal{H}} R_N(h) \tag{2.4}$$

This method is known as the *Empirical Risk Minimisation* inductive principle (ERM) Vapnik (1982; 1992). The ERM principle was thoroughly studied during the 1960–1990s by Vladimir Vapnik and Alexey Chervonenkis, as well as other scientists, and it is summarised in the book *The nature of statistical learning theory* Vapnik (1995). Empirical risk minimisation is a general principle and many classical estimation methods, such as least squares regression and maximum likelihood estimation, can be formulated as realisations of the ERM principle.

Some important aspects of the theory explained in the book and studied in a number of publications are the necessary and sufficient conditions for the consistency of algorithms based on the ERM, that is under what conditions minimising the empirical risk converges in the minimisation of the risk, when the number of examples tends to infinity Vapnik & Chervonenkis (1991); the rate of convergence of learning processes based on the ERM. Here, we will summarise the most important aspects and concepts of the learning theory based on the ERM.

Uniform bounds

Having outlined the notion of empirical risk minimisation and its components, we can define generalisation as the ability of a learning algorithm to achieve small risk $R(h)$. We can now return to the question of the feasibility of generalisation or the consistency of a learning process. An important result from probability theory that serves as starting point to study the feasibility of learning is Hoeffding's inequality Hoeffding (1963). It states that for any $\varepsilon > 0$ and a set of N i.i.d. random variables $Z_1 \dots Z_N$, such that $\mathbb{P}\left[a \leq Z_i \leq b\right] \forall i$, then:

$$\mathbb{P}\left[\left|\frac{1}{N}\sum_{i=1}^{N} Z_i - \mathbb{E}\left[Z\right]\right| > \varepsilon\right] \leq 2\exp\left(-\frac{2\varepsilon^2 N}{(b-a)^2}\right) \tag{2.5}$$

Hoeffding's inequality can be regarded as quantitative version of the law of large numbers[2] for the case when the variables are bounded. Equation 2.5 can be developed into a more useful form for our purposes, relating the risk of a hypothesis, $R(h)$, and the empirical risk, $R_N(h)$. For any $\delta > 0$, at least with probability $1 - \delta$:

$$R(h) \leq R_N(h) + \sqrt{\frac{1}{2N}\log\frac{2}{\delta}} \tag{2.6}$$

The interpretation of Equation 2.6 is that the risk of hypothesis h is bounded by a quantity that depends linearly and on the empirical risk plus a constant that depends on the number of samples used to obtain the empirical risk and the confidence δ. On the one hand, this is good news as it establishes that the empirical risk is indicative of the true risk when the number of samples is large. This confirms the feasibility of learning.

Nonetheless, on the other hand, the bound in Equation 2.6 is highly limited and useless in practice. Essentially, it says that for each hypothesis h, there exists a set of samples for which the bound holds. However, the function which will be chosen by the learning algorithm is unknown

[2]The law of large numbers is a fundamental theorem from probability theory. It states that the sample average converges in probability towards the expected value as the sample size increases.

before training it on the data set. For instance, there exists, as well, a function for which the empirical risk is not indicative of the true risk at all. In order to derive tighter, more useful bounds we need to take into account all the possible hypotheses $\mathcal{H}$ that the learning algorithm may choose. One of the best studied approaches is to consider *uniform bounds*.

The idea is to find an upper bound of the *supremum* of $R(h) - R_N(h)$, as that will clearly provide an upper bound on $R(g)$:

$$R(g) - R_N(g) \leq \sup_{h \in \mathcal{H}} [R(h) - R_N(h)] \tag{2.7}$$

The simplest way is to consider the disjunction of all events $|R(h_m) - R_N(h_m)| > \varepsilon$ for a finite set of hypothesis $\mathcal{H}$ of size M and $m = 1, \ldots, M$. Then, by applying the union bound[3] to the application of Hoeffding's inequality (see Equation 2.5) to the union of the difference of the risk and the empirical risk for all hypotheses, we get that

$$\begin{aligned} \mathbb{P}\left[|R(h) - R_N(h)| > \varepsilon\right] &\leq \sum_{m=1}^{M} \mathbb{P}\left[|R(h_m) - R_N(h_m)| > \varepsilon\right] \\ &\leq 2M \exp(-2\varepsilon^2 N) \end{aligned} \tag{2.8}$$

and hence, equivalent to Equation 2.6, for $\delta > 0$, with probability at least $1 - \delta$:

$$R(g) \leq R_N(g) + \sqrt{\frac{1}{2N} \log \frac{2M}{\delta}} \tag{2.9}$$

This is finally a practical error bound that guarantees that the risk of the final hypothesis g found via empirical risk minimisation will be bounded by the empirical risk measured on the training set, as long as the hypothesis set is finite, that is $|\mathcal{H}| = M$. A subtler implication of Equation 2.8, besides the fact that it leads to Equation 2.9, is that $R(g) \geq R_N(g) - \varepsilon$ also holds. Hence, the ERM principle ensures that there are not much better hypotheses than g in the set $\mathcal{H}$.

While the generalisation bound based on uniform deviations is a key result in statistical learning theory and it is theoretically relevant, it is only valid for learning algorithms that operate with finite hypothesis sets. For many machine learning algorithms, the hypothesis sets are infinitely large and additional theoretical results are necessary to describe their generalisation guarantees. Below we present some of the most important results.

Vapnik-Chervonenkis theory

When the hypothesis set $\mathcal{H}$ is uncountable and thus the right hand side of Equation 2.9 is unbounded, a classical approach is to use the notion of *growth function*, also known as *shatter coefficient* or *shattering number*. The growth function $m_{\mathcal{H}}(N)$ was introduced by Vapnik & Chervonenkis (1971) and it denotes the maximal *effective* size of $\mathcal{H}$ on a set of N examples, that is the maximum number of ways into which N data points can be classified by the function class. Formally:

$$m_{\mathcal{H}}(N) = \sup_{x_1, \ldots, x_N \in \mathcal{X}} |(h(x_1), \ldots, h(x_N)) : h \in \mathcal{H}| \tag{2.10}$$

For the case of binary classification that we are considering, the growth function $m_{\mathcal{H}}(N) \leq 2^N$. If the hypothesis set is capable of generating all possible dichotomies (binary labellings) of $x_1, \ldots, x_N$, then $\mathcal{H}$ is said to *shatter* the data set. If no data set of size k can be shattered by $\mathcal{H}$, then k is said to be a *break point* for $\mathcal{H}$.

One important concept in statistical learning theory is the Vapnik-Chervonenkis dimension, known as *VC dimension* for short. The VC dimension of a hypothesis set $\mathcal{H}$, denoted by $d_{VC}(\mathcal{H})$

[3] $\mathbb{P}(\bigcup_I A_i) \leq \mathbb{P}(\sum_i A_i)$

or simply d_{VC} is the largest N such that $m_{\mathcal{H}}(N) = 2^N$, that is the largest data set size that the hypothesis can shatter. Hence, if d_{VC} is the VC dimension of $\mathcal{H}$, then $k = d_{VC} + 1$ is a break point for the growth function.

It can be shown that if a hypothesis class $\mathcal{H}$ has finite VC dimension d_{VC}, then the growth function can be upper bounded by a polynomial:

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{d_{VC}} \binom{N}{i} \tag{2.11}$$

This allows us to bound the risk of a hypothesis in terms of the empirical risk and the growth function, what is known as the *VC generalisation bound*. For $\delta > 0$, with probability at least $1 - \delta$:

$$R(g) \leq R_N(g) + \sqrt{\frac{8}{N} \log\left(\frac{4m_{\mathcal{H}}(2N)}{\delta}\right)} \tag{2.12}$$

The VC generalisation bound is a key result in statistical learning theory as it establishes the feasibility of learning with infinite hypothesis sets: with enough data, all hypotheses in an infinite $\mathcal{H}$ with finite VC dimension will generalise from the empirical risk. The bound holds for all hypothesis sets, learning algorithms, input spaces, probability distributions and binary targets (Abu-Mostafa et al., 2012). Such generality comes at the expense of being quite a loose bound to be used in practice.

One interpretation of the generalisation bound in Equation 2.12 that we will use in this thesis is that the right hand side consists of two terms: the empirical risk and a term that is usually interpreted as a penalty for model complexity:

$$\Omega(N, \delta, \mathcal{H}) = \sqrt{\frac{8}{N} \log\left(\frac{4m_{\mathcal{H}}(2N)}{\delta}\right)} \leq \sqrt{\frac{8}{N} \log\left(\frac{4(2N)^{d_{VC}}}{\delta}\right)} \tag{2.13}$$

Ω depends on the number of examples N, the confidence parameter δ and the hypothesis class $\mathcal{H}$. The bound gets tighter (better) as the number of examples increases, as the confidence constant δ increases and as the complexity of the hypothesis set decreases (lower d_{VC}). This form of the bound on the risk, $R(g) \leq R_N(g) + \Omega(N, \delta, \mathcal{H})$, is found in most methods to estimate the theoretical generalisation guarantees of learning algorithms.

Rademacher complexity

Given these limitations of the Vapnik-Chervonenkis theory and, in particular, of the VC dimension, other measures of complexity have been developed (Bartlett et al., 2002). One relatively recent, popular example is the *Rademacher complexity*, which allows to define generalisation bounds that are not restricted to binary classification and hold for any class of real-valued functions.

Let $\sigma_1, \ldots, \sigma_N$ be a set of independent random variables such that $P(\sigma_i = 1) = P(\sigma_i = -1) = \frac{1}{2}$. These are known as *Rademacher variables*, hence the name of the complexity measure. As before, we consider a sample of N independent data points $x_1, \ldots, x_N$ defined on $\mathcal{X}$. Now, instead of being restricted to binary classification, we let $\mathcal{F}$ be the class of real-valued functions $f \colon \mathcal{X} \mapsto \mathbb{R}$. Then, the *empirical Rademacher complexity* of $\mathcal{F}$ with respect to the sample of size N is defined as:

$$\hat{\mathcal{R}}_N(\mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{N} \sum_{i=1}^{N} \sigma_i f(x_i) \right| \right] \tag{2.14}$$

where $\mathbb{E}_\sigma$ denotes the expectation with respect to the Rademacher variables. The *Rademacher complexity*, also found in the literature as *Rademacher average*, is defined as the expectation of the empirical Rademacher complexity over all data sets of size N on $\mathcal{X}$:

$$\mathcal{R}_N(\mathcal{F}) = \mathbb{E}\left[\hat{\mathcal{R}}_N(\mathcal{F})\right] \tag{2.15}$$

The interpretation of the Rademacher average as a complexity measure is intuitive: It is a measure of the ability of the function class $\mathcal{F}$ to fit random noise, introduced by the Rademacher variables σ_i. For a very large and complex $\mathcal{F}$, there will be a function f that can fit the noise, making $\hat{\mathcal{R}}_N(\mathcal{F})$ larger.

For the case of binary classification that we have considered so far, in which $\mathcal{H} \subseteq \{h\colon \mathcal{X} \mapsto \mathcal{Y} = \{-1, 1\}\}$, it can be easily shown that $\hat{\mathcal{R}}_N(\mathcal{H}) = \frac{1}{2}\hat{\mathcal{R}}_N(\mathcal{F})$, using the fact that σ_i and $\sigma_i Y_i$ have the same distribution. Finally, we can use the Rademacher complexity to bound the risk of the final hypothesis. For $\delta > 0$, with probability at least $1 - \delta$:

$$R(g) \leq R_N(g) + \hat{\mathcal{R}}_N + \sqrt{\frac{2\log\frac{2}{\delta}}{N}} \tag{2.16}$$

2.1.3 Regularisation

In Section 2.1.2, we have seen that the goal of a machine learning algorithm is to find a hypothesis $h(\mathbf{x})$ that, given some data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_i^N$, minimises the risk functional $R(h)$, which in turn depends on a loss function $L(y, h(\mathbf{x}))$ chosen as a criterion for the optimisation problem. We have also seen that, since it is not possible to exactly calculate the risk, in practice we optimise the empirical risk $R_N(h)$, which is calculated on the available training data:

$$g = \arg\min_{h \in \mathcal{H}} R_N(h) \tag{2.17}$$

This method is known as empirical risk minimisation (ERM), and in Section 2.1.2 we have summarised the theory that describes the convergence of the empirical risk to the actual risk of the model. Nonetheless, despite the importance of this theory to confirm the feasibility of learning, the bounds on the generalisation error are not always applicable in practice: they are quite loose, depend on hypothesis sets with finite VC dimension and, in general, the plain ERM principle is intended to deal with large sample sizes. In practice, learning algorithms rely on extensions of ERM, such as the principle of structural risk minimisation (SRM) (Vapnik & Chervonenkis, 1974). A particular case of SRM is regularisation, a widely used technique in machine learning and one of its cornerstones (Poggio & Girosi, 1990; Girosi et al., 1995). In this section we review the fundamentals of regularisation and present some of its most common forms, which are relevant for this thesis.

The concept of regularisation of learning algorithms is closely related to the mathematical problem of approximating a function from sparse data, that is finding $f \in \mathcal{F}$ such that $Af = F$. Hadamard (1902) demonstrated that under some general circumstances this is an ill-posed problem. That is, an arbitrarily small deviation ε of F (F_ε instead of F, where $\|F - F_\varepsilon\| < \varepsilon$) can cause large deviations in the solution of the equation. Formally, minimising the functional

$$\rho(f) = \|Af - F_\varepsilon\|^2 \tag{2.18}$$

is not guaranteed to provide a good approximation even if ε tends to zero. This closely resembles the learning problem that we have described above, where the task is to find the function g that best approximates the data $\mathcal{D}$, using the empirical risk, as summarised in Equation 2.17 and detailed in Section 2.1.2. As a matter of fact, finding g in the presence of noise is also ill-posed, as there is an infinite number of solutions. In order to find a suitable solution with access to only limited data, it is necessary to constrain the hypothesis space $\mathcal{H}$ with some a priori information, for instance assuming that the function is smooth. This is the idea of the regularisation principles discovered in the 1960s (Phillips, 1962; Tikhonov, 1963; Ivanov, 1976). In particular, they found that if instead of minimising the functional $\rho(f)$ of Equation 2.18, one minimises the so-called regularised functional

$$\rho^*(f) = \|Af - F_\varepsilon\|^2 + \lambda(\varepsilon)\Omega(f) \tag{2.19}$$

where $\Omega(f)$ is the regularisation functional, and $\lambda(\varepsilon)$ is a constant that determines the level of noise, then the sequence of solutions converges as $\varepsilon \to 0$. In our particular case of learning from

data, the principles of regularisation translate into adding a similar regularisation term to the objective function. The similarity between the two problems is most obvious if we consider the mean squared error loss, instead of binary classification. In this case, the optimisation problem becomes the following:

$$\begin{aligned} g &= \arg\min_{h\in\mathcal{H}} R_{reg}(h) = \arg\min_{h\in\mathcal{H}} \left[R_N(h) + \lambda\Omega(h)\right] \\ &= \arg\min_{h\in\mathcal{H}} \left[\sum_{i=1}^{N}(h(\mathbf{x}_i) - y_i)^2 + \lambda\Omega(h)\right] \end{aligned} \tag{2.20}$$

$\Omega(h)$ is the regularisation functional or regulariser, which incorporates prior information or desired properties of the model. In general, the regulariser is chosen to encourage smooth functions. The constant λ is the regularisation parameter, which controls the strength of the regularisation.

The concepts of generalisation and particularly regularisation are closely related to the widely used concept of *overfitting*, that is the tendency of a learning algorithm to excessively fit the training data points, to the detriment of its generalisation. Broadly speaking, a complex hypothesis function is more likely to *overfit* the training data than a simpler function. The idea of function smoothness introduced by the regularisation term can be seen as a way to counteract overfitting, in favour of better generalisation. This establishes a trade-off where ideally the learning algorithm should strike the right balance between fitting the data, that is minimising the empirical risk, and finding a smooth enough function that generalises well. This trade-off can be controlled by the value of the regularisation parameter, which is often determined through *cross-validation* (Stone, 1974; Allen, 1974).

The choice of $\Omega(h)$ leads to different forms of regularisation and there is a very large body of literature on this topic. Popular choices are constraints on the norm of the parameters, which we will discuss in Section 2.1.3 or constraints on the curvature of h. In modern machine learning, the concept of regularisation is very broad and regularisation is considered to be any mechanism that prevents overfitting, hence improving generalisation. In Chapter 3 we compare different forms of regularisation and discuss the distinction between implicit and explicit regularisation—key in Chapter 4—and other regularisation taxonomies.

In the remaining of this section we introduce two specific regularisation techniques, weight decay and dropout, which are arguably the two most common forms of regularisation in modern neural networks.

Weight decay

Weight decay is the common name used to refer to *L^2–norm regularisation*[4], which is in turn a particular case of L^p–norm regularisation. In this section we will first review L^p–norm regularisation as a direct realisation of the type of regularisation described above, and then present the specific aspects of weight decay.

In the previous section we have seen that the concept of regularisation derived from the mathematical tool for solving ill-posed problems, results in a modification of the objective function (see Equation 2.20). We will denote the regularised objective by $\hat{J}$:

$$\hat{J}(\boldsymbol{\theta}; \mathbf{x}, y) = J(\boldsymbol{\theta}; \mathbf{x}, y) + \lambda\Omega(\boldsymbol{\theta}) \tag{2.21}$$

L^p–norm regularisation refers to the family of techniques which apply a penalty on the norm

[4]Note, however, that in part of the machine literature, the term weight decay refers to a form of regularisation in which the L^2–norm penalty is added directly to the update rule of gradient descent. This results in a conceptually equivalent form of regularisation, but with a slight numerical difference. See Babenko (2018) for more details.

of the parameters:

$$\Omega(\boldsymbol{\theta}) = \phi(\|\boldsymbol{\theta}\|_p) = \phi\left(\left(\sum_{i=1}^{d}|\theta_i|^p\right)^{\frac{1}{p}}\right) \tag{2.22}$$

where $\phi(\cdot)$ is an optional function applied on the norm, for example the squared function. The most commonly used L^p–norm penalties are L^1 and L^2 regularisation. L^2 regularisation is probably one of the most widely used regularisation techniques in machine learning. In deep learning, it is commonly referred to as *weight decay*, but it is also known as Tikhonov regularisation (Tikhonov, 1963) or ridge regression when applied to linear regression. In this thesis, we will analyse weight decay in neural networks and compare it to other regularisation techniques in Chapter 4, and we use it to regularise a logistic regression algorithm in Chapter 7.

The specific regularisation term typically used for weight decay is $\Omega(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{\theta}\|_2^2$ because it allows to implement and express the objective function in a convenient and efficient way using the dot product between the vector of parameters and its transpose:

$$\hat{J}(\boldsymbol{\theta}; \mathbf{x}, y) = J(\boldsymbol{\theta}; \mathbf{x}, y) + \frac{\lambda}{2}\boldsymbol{\theta}^T\boldsymbol{\theta} \tag{2.23}$$

Weight decay has been long used, at least since the 1980s (Hinton, 1987), and widely studied, both empirically (Zhang et al., 2018) and theoretically (Krogh & Hertz, 1992; Neyshabur et al., 2015), especially in the context of neural networks. Intuitively, the mechanism provided by weight decay is to restrict the norm of the trainable parameters, by decreasing the weight vector at every iteration of a model trained with gradient descent, in the directions that do not contribute much to reducing the objective function. Relevant to this dissertation is the result by Bishop (1995), which showed that optimising a squared error loss with weight decay is equivalent to training with random noise in the inputs. In Chapter 4, we will use this result to derive some theoretical insights into the comparison of weight decay and data augmentation.

Dropout

Dropout is a regularisation technique first described by this name in (Hinton et al., 2012; Srivastava et al., 2014), although closely related to *dilution* (Hertz et al., 1991). It is very widely used in modern neural networks due to its simplicity and effectiveness. In practice, dropout is implemented and hence can be described as a method that omits every unit—parameter, feature detector, etc.—of a model with probability p, at every iteration of the optimisation process (training). At inference (test) time, the whole set of units is considered. While dropout can be applied to a broad class of models, it is most often used to train deep neural networks and for simplicity we will also consider neural networks in this section.

Dropout is often described as a practical approximation of training an ensemble of models through bootstrap aggregation, commonly known as *bagging* (Breiman, 1996), in which M models are trained on M subsets of a data set of size N uniformly sampled with replacement (bootstrap sample). At inference time, the outputs of the M models on each data point are averaged (for regression) or combined through majority voting (for classification). Bagging is a widely used technique, known to reduce the variance and overfitting of learning algorithms. However, it is computationally expensive as it requires training multiple models. Dropout efficiently approximates a form of bagging with an exponentially large number of sub-networks (models). Since neural networks are typically trained with mini-batch iterative methods (such as stochastic gradient descent), the parameters of the model are updated by computing the loss of a sub-network on a sub-sample of the data set. An important difference between standard bagging and dropout is that while in bagging the models are independent, with dropout the models share a subset of the parameters from the parent neural network.

In (Srivastava et al., 2014), dropout training is connected with a theory by Livnat et al. (2010) about the superiority of sexual over asexual reproduction in nature. According to this theory, a

criterion for natural selection would be enhancing the robust combination of different genes for better adaptation to changes, as opposed to the optimisation of the individual fitness through a slight mutation of one parent's genes. Sexual reproduction would favour this criterion by preventing co-adaptations of the available genes in one individual. With dropout, the units of a network are forced to learn useful combinations with other subsets of random units, hence preventing co-adaptation and increasing robustness.

Dropout has greatly impacted the deep learning community[5]. It is widely used for training neural networks in both research and application and it has been deeply studied both empirically and theoretically (Gal & Ghahramani, 2016), sometimes uncovering contradictory and surprising properties. While it is out of the scope to review the vast literature on dropout training, we can mention some relevant findings. Since shortly after it was proposed, dropout has been analysed as adaptive form of regularisation (Wager et al., 2013). Baldi & Sadowski (2013) found that the dynamics of gradient descent with dropout training approximate that of a regularised error function, while Helmbold & Long (2017) showed that in deeper networks, the behaviour of dropout differs significantly from standard regularisation. More recently, Mou et al. (2018) derived generalisation bounds based on the Rademacher complexity for deep neural networks trained with dropout. Finally, an interesting and relevant finding for this thesis is that dropout applied to the intermediate units of a neural network has been shown to be equivalent to training with noise in the input (Bouthillier et al., 2015).

[5]The two original papers have been increasingly cited almost 25,000 times at the time of writing, according to Google Scholar.

Bibliography

Abu-Mostafa, Y. S., Magdon-Ismail, M., and Lin, H.-T. *Learning from data*. AMLBooks, 2012.

Allen, D. M. The relationship between variable selection and data agumentation and a method for prediction. *Technometrics*, 1974.

Alpaydin, E. *Introduction to machine learning*. MIT Press, 2009.

Babenko, B. weight decay vs L2 regularization. Accessed: 2020-05-20, 2018. URL `https://bbabenko.github.io/weight-decay/`.

Baldi, P. and Sadowski, P. J. Understanding dropout. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

Bartlett, P. L., Boucheron, S., and Lugosi, G. Model selection and error estimation. *Machine Learning*, 2002.

Bishop, C. M. Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, 1995.

Bousquet, O., Boucheron, S., and Lugosi, G. Introduction to statistical learning theory. In *Advanced Lectures on Machine Learning*. 2003.

Bouthillier, X., Konda, K., Vincent, P., and Memisevic, R. Dropout as data augmentation. *arXiv preprint arXiv:1506.08700*, 2015.

Breiman, L. Bagging predictors. *Machine Learning*, 1996.

Gal, Y. and Ghahramani, Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*, 2016.

Girosi, F., Jones, M., and Poggio, T. Regularization theory and neural networks architectures. *Neural Computation*, 1995.

Hadamard, J. Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton University Bulletin*, 1902.

Helmbold, D. P. and Long, P. M. Surprising properties of dropout in deep networks. *Journal of Machine Learning Research (JMLR)*, 2017.

Hertz, J., Krogh, A., and Palmer, R. G. *Introduction to the Theory of Neural Computation*. Addison-Wesley Longman Publishing Co., Inc., USA, 1991.

Hinton, G. E. Learning translation invariant recognition in a massively parallel networks. In *International Conference on Parallel Architectures and Languages Europe*. 1987.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 1963.

Ivanov, V. V. *The theory of approximate methods and their applications to the numerical solution of singular integral equations*. Nordhof International, 1976.

Krogh, A. and Hertz, J. A. A simple weight decay can improve generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1992.

Livnat, A., Papadimitriou, C., Pippenger, N., and Feldman, M. W. Sex, mixability, and modularity. *Proceedings of the National Academy of Sciences (PNAS)*, 2010.

Mou, W., Zhou, Y., Gao, J., and Wang, L. Dropout training, data-dependent regularization, and generalization bounds. In *International Conference on Machine Learning (ICML)*, 2018.

Murphy, K. P. *Machine learning: a probabilistic perspective*. MIT Press, 2012.

Neyshabur, B., Tomioka, R., and Srebro, N. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, 2015.

Phillips, D. L. A technique for the numerical solution of certain integral equations of the first kind. *Association for Computing Machinery*, 1962.

Poggio, T. and Girosi, F. Networks for approximation and learning. *Proceedings of the IEEE*, 1990.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research (JMLR)*, 2014.

Stone, M. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1974.

Tikhonov, A. N. On solving ill-posed problem and method of regularization. *Doklady Akademii Nauk USSR*, 1963.

Vapnik, V. N. *Estimation of dependences based on empirical data*. Springer, 1982.

Vapnik, V. N. Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1992.

Vapnik, V. N. *The nature of statistical learning theory*. Springer Verlag, 1995.

Vapnik, V. N. and Chervonenkis, A. Y. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 1971.

Vapnik, V. N. and Chervonenkis, A. Y. Theory of pattern recognition. 1974.

Vapnik, V. N. and Chervonenkis, A. Y. The necessary and sufficient conditions for consistency of the method of empirical risk minimization. *Pattern Recognition and Image Analysis*, 1991.

Von Luxburg, U. and Schölkopf, B. Statistical learning theory: Models, concepts, and results. In *Handbook of the History of Logic*, volume 10, pp. 651–706. Elsevier, 2011.

Wager, S., Wang, S., and Liang, P. S. Dropout training as adaptive regularization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

Zhang, G., Wang, C., Xu, B., and Grosse, R. Three mechanisms of weight decay regularization. *arXiv preprint arXiv:1810.12281*, 2018.

Chapter 3

Explicit and implicit regularisation

Outreach

This chapter extends the following publications:

- *Data augmentation instead of explicit regularization.* Alex Hernández-García, Peter König. arXiv preprint arXiv:1806.03852, 2018.

One of the central issues in machine learning research and application is finding ways of improving generalisation. Regularisation, broadly defined as any modification applied to a learning algorithm that helps the model generalise better, plays therefore a key role in machine learning[1]. In the case of deep learning, where neural networks tend to have several orders of magnitude more parameters than training examples, statistical learning theory (Section 2.1.2) indicates that regularisation becomes even more crucial. Accordingly, a myriad of techniques have been proposed as regularisers (Section 2.1.3): weight decay (Hanson & Pratt, 1989) and other L^p penalties on the learnable parameters; dropout—random dropping of units during training—(Srivastava et al., 2014) and stochastic depth—random dropping of whole layers—(Huang et al., 2016), to name a few.

Moreover, whereas in simpler machine learning algorithms the regularisers can be easily identified as explicit terms in the objective function, in modern deep neural networks the sources of regularisation are not only explicit but implicit (Neyshabur et al., 2014). In this regard, many techniques have been studied for their regularisation effect, despite not being explicitly intended as such. Examples are convolutional layers (LeCun et al., 1990), batch normalisation (Ioffe & Szegedy, 2015) and data augmentation. In sum, there are multiple elements in deep learning that contribute to reducing overfitting and thus improve generalisation.

It is common practice in both the scientific literature and application to incorporate several of these regularisation techniques in the training procedure of neural networks. For instance, weight decay, dropout and data augmentation have been used jointly in multiple well-known architectures (Tan & Le, 2019; Huang et al., 2017; Zagoruyko & Komodakis, 2016; Springenberg et al., 2014). It is therefore implicitly assumed that each technique is necessary and contributes additively to improving generalisation. However, the interplay between regularisation techniques is yet to be well understood and might be an important piece for the puzzle of why and when deep networks generalise.

In Chapter 4, we will focus on contrasting some specific forms of regularisation, namely weight decay, dropout and data augmentation. This chapter serves as a preamble of the following one.

[1] In Chapter 2 we review the fundamentals of machine learning and in particular, Section 2.1.3 reviews the essential aspects of regularisation to understand this and the upcoming chapters.

Here, we will provide definitions of two terms that have been widely but ambiguously used in the machine learning literature: explicit and implicit regularisation. We contend that these terms are useful to understand and explain the role of regularisation in artificial neural networks, as reflected by their use in the literature. Therefore, it is important to settle the precise meaning of the terms and provide examples. Besides being helpful to interpret the results in Chapter 4, we also hope that this is a useful contribution to the machine learning community.

3.1 Why do we need definitions?

While several regularisation taxonomies have been proposed (see Section 3.3), to the best of our knowledge there is no formal definitions of explicit and implicit regularisation in the machine literature. Nonetheless, the terms have been widely used Neyshabur et al. (2014); Zhang et al. (2017a); Wilson et al. (2017); Mesnil et al. (2011); Poggio et al. (2017); Martin & Mahoney (2018); Achille & Soatto (2018). This could suggest that the concepts are ingrained in the field and well understood by the community. However, by analysing the use of the terms explicit and implicit regularisation in the literature and in discussions with practitioners one can see that there is a high degree of ambiguity. In this section we will review some examples and motivate the need for formal definitions.

The PhD thesis by Neyshabur (2017) is devoted to the study of implicit regularisation in deep learning. For instance, Neyshabur shows that common optimisation methods for deep learning, such as stochastic gradient descent (SGD), introduce an inductive bias that lead to better generalisation. That is, SGD *implicitly* regularises the learning process. However, the notion and definition of implicit regularisation is only implied in Neyshabur's PhD thesis and related works.

Some may argue that the definitions are not necessary. By looking at one single piece of work, even without a formal definition, the meaning may be inferred from the use. However, when one considers a larger body of work by multiple authors, differences and even contradictions emerge. In the work by Neyshabur and colleagues (Neyshabur, 2017; Neyshabur et al., 2014), it can be interpreted that implicit regularisation refers to the generalisation improvement provided by techniques such as stochastic gradient descent (SGD) that are not *typically* considered as regularisation. By extension, explicit regularisation would refer to those other techniques: "we are not including any explicit regularisation, neither as an explicit penalty term nor by modifying optimisation through, e.g., drop-outs, weight decay, or with one-pass stochastic methods" (Neyshabur, 2017). In Poggio et al. (2017), it can be interpreted that implicit regularisation refers to techniques that lead to minimisation of the parameter norm without explicitly optimising for it. By extension, explicit regularisation would refer to classical penalties on the parameter norm, such as weight decay. It is therefore unclear whether other methods such as dropout should be considered explicit or implicit regularisation according to Poggio et al. (2017).

Zhang et al. (2017a) raised the thought-provoking idea that "explicit regularisation may improve generalisation performance, but is neither necessary nor by itself sufficient for controlling generalisation error." The authors came to this conclusion from the observation that turning off the "explicit regularisers" of a model does not prevent the model from generalising reasonably well. In their experiments, the explicit regularisation techniques they turned off were, specifically, weight decay, dropout and data augmentation. In this case, it seems that Zhang et al. (2017a) made a distinction based on the mere intention of the practitioner. Under that logic, because data augmentation has to be designed and applied *explicitly*, it would be explicit regularisation.

These examples illustrate that the terms explicit and implicit regularisation have been used subjectively and inconsistently in the literature. In order to help avoid ambiguity, settle the concepts and facilitate the discussion, in the next section we propose definitions and provide examples to illustrate each category. Further, we will argue that data augmentation is not explicit regularisation and introduce some key differences with respect to explicit regularisation, which will set the grounds for Chapter 4.

3.2 Definitions and examples

We propose the following definitions of explicit and implicit regularisation:

- **Explicit regularisation techniques** are those techniques which reduce the *representational* capacity of the model class they are applied on. That is, given a model class $\mathcal{H}_0$, for instance a neural network architecture, the introduction of explicit regularisation will span a new hypothesis set $\mathcal{H}_1$, which is a *proper subset* of the original set, that is $\mathcal{H}_1 \subsetneq \mathcal{H}_0$.
- **Implicit regularisation** is the reduction of the generalisation error or overfitting provided by means other than explicit regularisation techniques. Elements that provide implicit regularisation do not reduce the *representational* capacity, but may affect the *effective* capacity of the model: the *achievable* set of hypotheses given the model, the optimisation algorithm, hyperparameters, etc.

Note that we define explicit and implicit regularisation by using the concepts of *representational* and *effective* capacity. Although these terms are also used ambiguously by some practitioners, definitions of these concepts can be found in the literature. For instance, the textbook Deep Learning (Goodfellow et al., 2016) clearly defines the representational capacity as the "the family of functions the learning algorithm can choose from" and explains that the effective capacity "may be less than the representational capacity" because the learning algorithm does not always find the "best function" due to "limitations, such as the imperfection of the optimisation algorithm". Thinking of these *limitations* as implicit regularisation denotes that this can be beneficial. In any case, we here adopt these definitions of representational and effective capacity.

One of the most common explicit regularisation techniques in machine learning is L^p-norm regularisation (Tikhonov, 1963), of which weight decay is a particular case, widely used in deep learning. Weight decay sets a penalty on the L^2 norm of the model's learnable parameters, thus constraining the representational capacity of the model. Dropout (Srivastava et al., 2014) is another common example of explicit regularisation, where the hypothesis set is reduced by stochastically deactivating a number of neurons during training. Similar to dropout, stochastic depth (Huang et al., 2016), which drops whole layers instead of neurons, is also an explicit regularisation technique.

Regarding implicit regularisation, note first that the above definition does not refer to *techniques*—as in the definition of explicit regularisation—but to a regularisation *effect*, as it can be provided by multiple elements of different nature. For instance, stochastic gradient descent (SGD) is known to have an implicit regularisation effect—reduction of the generalisation error—without constraining the representational capacity (Zhang et al., 2017b). Batch normalisation neither reduces the capacity, but it improves generalisation by smoothing the optimisation landscape (Santurkar et al., 2018). Of quite a different nature, but still implicit, is the regularisation effect provided by early stopping (Yao et al., 2007), which does not reduce the representational but the effective capacity.

In these examples and all other cases of implicit regularisation, we can think of the effect on the capacity in the following way: we start by defining our model class, for instance a neural network, which spans a set of functions $\mathcal{H}_0$ (see Section 2.1.1). If we decide to train with explicit regularisation, for instance weight decay or dropout, then the model will have access to a smaller set of functions $\mathcal{H}_1 \subsetneq \mathcal{H}_0$, that is the representational capacity. On the contrary, if we decide to train with SGD, batch normalisation or early stopping, the set of functions spanned by the model stays identical. Due to the dynamics and limitations imposed by these techniques, some functions may never be found, but theoretically they could be. In other words, the effective capacity may be smaller but not the representational capacity.

Central to this thesis is data augmentation, a technique that provides an implicit regularisation effect. As we have discussed, Zhang et al. (2017a) considered data augmentation an explicit regularisation technique and was analysed as equivalent in terms of category to weight decay and dropout. However, data augmentation does not reduce the representational capacity of the models and hence, according to our definitions, cannot be considered explicit regularisation. This

is relevant to understand the differences between weight decay, dropout and data augmentation that we will present in Chapter 4, especially in the context of artificial neural networks.

3.3 On the taxonomy of regularisation

As in most disciplines, many taxonomies of regularisation techniques for machine learning have been proposed. Being a key ingredient of machine learning theory and practice, machine learning textbooks include a review of regularisation methods. In the case of deep learning, besides the classical regularisation methods used in *traditional* machine learning, multiple new regularisation techniques have been proposed in recent years, and many techniques have been analysed because of their implicit regularisation effect. In this section, we briefly review some taxonomies of regularisation proposed in the literature and discuss their similarity with our definitions.

In their textbook, Goodfellow et al. (2016) review some of the most common regularisation techniques used to train deep neural networks, but do not discuss the concepts of explicit and implicit regularisation. More recently, Kukačka et al. (2017) provided an extensive review of regularisation methods for deep learning. Although they mention the implicit regularisation effect of techniques such as SGD, no further discussion of the concepts is provided. Nonetheless, they define the category *regularisation via optimisation*, which is somewhat related to implicit regularisation. However, regularisation via optimisation is more specific than our definition; hence, methods such as data augmentation would not fall into that category.

Recently, Guo et al. (2019) provided a distinction between *data-independent* and *-dependent* regularisation. They define data-independent regularisation as those techniques that impose certain constraint on the hypothesis set, thus constraining the optimisation problem. Examples are weight decay and dropout. We believe this is closely related to our definition of explicit regularisation. On the other hand, they define data-dependent regularisation as those techniques that make assumptions on the hypothesis set with respect to the training data, as is the case of data augmentation. While we acknowledge the usefulness of such taxonomy, we argue that the division between data-independent and -dependent regularisation leaves some ambiguity about other techniques, such as batch-normalisation, which neither imposes an explicit constraint on the representational capacity nor on the training data.

On the contrary, our distinction between explicit and implicit regularisation aims at being complete, since implicit regularisation refers to any regularisation effect that does not come from explicit—or data-independent—techniques.

3.4 Discussion

The main contribution of this chapter has been the proposal of definitions of explicit and implicit regularisation. These terms that have been widely used in the machine learning literature without being formally defined, hence giving rise to subjective and ambiguous use. With the definition of these important concepts we have set the grounds for our discussion on the rest of this thesis, especially in Chapter 4, but we also hope to help settle the concepts and reduce the ambiguity in the literature.

Besides this contribution, it is interesting to draw some connections between the concept of implicit regularisation, the discussion about inductive biases in the Introduction (Chapter 1) and data augmentation. According to our definition above, implicit regularisation is the improvement in generalisation provided by elements that are not explicit regularisation techniques. This is a broad definition that includes many possible sources of implicit regularisation. A concept that underlies many of them is that of inductive bias. The inductive bias encoded in explicit regularisation techniques is simply that smaller models generalise better, which is reminiscent of Occam's razor. While this is a powerful inductive bias, we have discussed that many other sources of inductive bias are possible and are worth exploring.

In this regard, a clear distinction between explicit and implicit regularisation may help in the analysis, especially in the case of deep learning. A striking difference between neural networks and other machine learning algorithms is that deep networks easily scale to an (almost) arbitrarily large number of parameters and still generalise well on a held out test data. Only recently are we starting to understand this phenomenon, which seemed to be at odds with the results of statistical learning theory (Belkin et al., 2019).

First, the concept of implicit regularisation may help explain the generalisation of deep neural networks. Second, the fact that very large neural networks can generalise well directly casts doubts on the need for explicit regularisation (Zhang et al., 2017a), that is to constrain the representational capacity. However, most artificial neural networks are still trained with explicit regularisation methods such as weight decay and dropout. In the next chapter, we follow up this idea and directly address the question of whether explicit regularisation is necessary in deep learning, provided enough implicit regularisation is included, specifically data augmentation.

Bibliography

Achille, A. and Soatto, S. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research (JMLR)*, 2018.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences (PNAS)*, 2019.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep learning*. MIT press, 2016.

Guo, H., Mao, Y., and Zhang, R. Mixup as locally linear out-of-manifold regularization. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2019.

Hanson, S. J. and Pratt, L. Y. Comparing biases for minimal network construction with back-propagation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1989.

Huang, G., Sun, Y., Liu, Z., Sedra, D., and Weinberger, K. Q. Deep networks with stochastic depth. In *European Conference on Computer Vision (ECCV)*. 2016.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, 2015.

Kukačka, J., Golkov, V., and Cremers, D. Regularization for deep learning: A taxonomy. *arXiv preprint arXiv:1710.10686*, 2017.

LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., and Jackel, L. D. Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1990.

Martin, C. H. and Mahoney, M. W. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *arXiv preprint arXiv:1810.01075*, 2018.

Mesnil, G. et al. Unsupervised and transfer learning challenge: a deep learning approach. In *Workshop on Unsupervised and Transfer Learning (ICML)*, 2011.

Neyshabur, B. *Implicit regularization in deep learning*. PhD thesis, Toyota Technological Institute at Chicago, arXiv:1709.01953, 2017.

Neyshabur, B., Tomioka, R., and Srebro, N. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *International Conference on Learning Representations (ICLR), arXiv:1412.6614*, 2014.

Poggio, T., Kawaguchi, K., Liao, Q., Miranda, B., Rosasco, L., Boix, X., Hidary, J., and Mhaskar, H. Theory of deep learning III: explaining the non-overfitting puzzle. *arXiv preprint arXiv:1801.00173*, 2017.

Santurkar, S., Tsipras, D., Ilyas, A., and Madry, A. How does batch normalization help optimization? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. Striving for simplicity: The all convolutional net. In *International Conference on Learning Representations (ICLR), arXiv:1412.6806*, 2014.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research (JMLR)*, 2014.

Tan, M. and Le, Q. V. EfficientNet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning (ICML)*, 2019.

Tikhonov, A. N. On solving ill-posed problem and method of regularization. *Doklady Akademii Nauk USSR*, 1963.

Wilson, A. C., Roelofs, R., Stern, M., Srebro, N., and Recht, B. The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Yao, Y., Rosasco, L., and Caponnetto, A. On early stopping in gradient descent learning. *Constructive Approximation*, 2007.

Zagoruyko, S. and Komodakis, N. Wide residual networks. In *British Machine Vision Conference (BMVC)*, 2016.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR), arXiv:1611.03530*, 2017a.

Zhang, C., Liao, Q., Rakhlin, A., Sridharan, K., Miranda, B., Golowich, N., and Poggio, T. Theory of deep learning III: Generalization properties of SGD. Technical report, Center for Brains, Minds and Machines (CBMM), 2017b.

Chapter 4

Data augmentation instead of explicit regularisation

Outreach

This chapter extends the following publications:

- *Data augmentation instead of explicit regularization.* **Alex Hernández-García**, Peter König. arXiv preprint arXiv:1806.03852, 2018.
- *Do deep nets really need weight decay and dropout?.* **Alex Hernández-García**, Peter König. arXiv preprint arXiv:1802.07042, 2018.
- *Further advantages of data augmentation on convolutional neural networks.* **Alex Hernández-García**, Peter König. International Conference on Artificial Neural Networks (ICANN, Best Paper Award), 2018.

Data augmentation in machine learning refers to the techniques that synthetically create new examples from a data set by applying possibly stochastic transformations on the existing examples. In the image domain, these transformations can be, for instance, slight translations or rotations, which preserve the perceptual appearance of the original images, but significantly alter the actual pixel values. Despite being an old technique (Abu-Mostafa, 1990; Simard et al., 1992) and ubiquitous in the deep learning literature and practice, data augmentation has often been regarded as a sort of *cheating*, *lower class* technique[1], which should not be used in order to assess the actual strength of a new proposal (Goodfellow et al., 2013; Graham, 2014; Larsson et al., 2016; Goodfellow et al., 2016). A common criticism is that data augmentation usually requires domain or expert knowledge and it cannot be easily generalised across data domains (DeVries & Taylor, 2017b).

Explicit regularisation methods such as weight decay (Hanson & Pratt, 1989) and dropout (Srivastava et al., 2014) are also nearly ubiquitous. In contrast, they are considered intrinsic parts of the learning algorithm and thus have remained unquestioned. However, in Chapter 3, we have introduced the differences between explicit and implicit regularisation and raised the question of

[1] As a result, the machine learning scientific community has heavily ignored data augmentation as a subject of study until recently. By way of illustration, the textbook Deep Learning (Goodfellow et al., 2016) dedicates one and a half pages to data augmentation, of which one third is devoted to the caveats of data augmentation. Only in the last few years has data augmentation started to receive increasing attention, likely due to the success of some data augmentation techniques, such as *cutout* (DeVries & Taylor, 2017a) and *mixup* (Zhang et al., 2017b), and especially by the popularisation by Google of *automatic* data augmentation (Cubuk et al., 2018), previously proposed by various university groups (Hauberg et al., 2016; Antoniou et al., 2017; Ratner et al., 2017; Lemley et al., 2017). We first submitted the results presented in this chapter in 2017 (Hernández-García & König, 2018) and other authors have also presented surveys on data augmentation techniques (Perez & Wang, 2017; Shorten & Khoshgoftaar, 2019). Promisingly, very recently has data augmentation started to be analysed as well from a theoretical point of view (Rajput et al., 2019; Chen et al., 2019; Lyle et al., 2020)

whether explicit regularisation is necessary in deep learning. On the other hand, in the Introduction (Chapter 1) we have discussed the view of data augmentation as a powerful inductive bias from visual perception. Building upong these insights, in this chapter, we analyse the role of data augmentation in neural networks trained for image object categorisation and the need for weight decay and dropout when data augmentation is used. We first derive some theoretical insights from statistical learning theory and then present the results of a large empirical study in which we contrast the contributions of each technique.

4.1 Theoretical insights

As we have reviewed in Section 2.1.2, the generalisation of a model class $\mathcal{H}$ can be analysed through complexity measures such as the VC-dimension or, more generally, the Rademacher complexity $\mathcal{R}_N(\mathcal{H}) = \mathbb{E}\left[\hat{\mathcal{R}}_N(\mathcal{H})\right]$, where:

$$\hat{\mathcal{R}}_N(\mathcal{H}) = \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}} \left| \frac{1}{N} \sum_{i=1}^{N} \sigma_i h(x_i) \right| \right] \tag{4.1}$$

is the empirical Rademacher complexity, defined with respect to a specific set of N data samples. Then, in the case of binary classification and the class of linear separators, the generalisation error of a hypothesis, $\hat{\epsilon}_N(h)$, can be bounded using the Rademacher complexity:

$$\hat{\epsilon}_N(h) \leq \mathcal{R}_N(\mathcal{H}) + \mathcal{O}\left(\sqrt{\frac{\ln 1/\delta}{N}}\right) \tag{4.2}$$

with probability $1 - \delta$. Tighter bounds for some model classes, such as fully connected neural networks, can be obtained (Bartlett & Mendelson, 2002), but it is not trivial to formally analyse the influence on generalisation of specific architectures or techniques. Nonetheless, we can use these theoretical insights to discuss the differences between explicit regularisation—specifically weight decay and dropout—and data augmentation.

A straightforward yet very relevant conclusion from the analysis of any generalisation bound is the strong dependence on the number of training examples N. Increasing N drastically improves the generalisation guarantees, as reflected by the second term in the right hand side of Equation 4.2 and by the dependence of the Rademacher complexity (Equation 4.1) on the sample size too. Data augmentation exploits prior knowledge about the data domain and aspects of visual perception—in the case of image object recognition—to create new examples and its impact on generalisation is related to an increment in N, as stochastic data augmentation can generate virtually infinite different samples. Admittedly, the augmented samples are not independent and identically distributed and thus, the effective increment of samples does not strictly correspond to the increment in N. This is why formally analysing the impact of data augmentation on generalisation is complex. Recent studies have made progress in this direction by analysing the effect of simple data transformations on generalisation from a theoretical point of view (Chen et al., 2019; Rajput et al., 2019).

Explicit regularisation methods aim, in contrast, at improving the generalisation error by constraining the hypothesis class $\mathcal{H}$ to reduce its complexity $\mathcal{R}_N(\mathcal{H})$ and, in turn, the generalisation error $\hat{\epsilon}_N(h)$. Crucially, while data augmentation exploits domain knowledge, most explicit regularisation methods only *naively* constrain the hypothesis class, by simply reducing the representational capacity, as we have discussed in the previous chapter. For instance, weight decay constrains the learnable models $\mathcal{H}$ by setting a penalty on the weights norm. Interestingly, Bartlett et al. (2017) showed that weight decay has little impact on the generalisation bounds and confidence margins. Dropout has been extensively used and studied as a regularisation method for neural networks (Wager et al., 2013), but the exact way in which it impacts generalisation is still an open question. In fact, it has been stated that the effect of dropout on neural networks

is "somewhat mysterious", complicated and its penalty highly non-convex (Helmbold & Long, 2017). Recently, Mou et al. (2018) have established new generalisation bounds on the variance induced by a specific type of dropout on feedforward networks.

An interesting observation is that dropout can be analysed as a random form of data augmentation without domain knowledge (Bouthillier et al., 2015). This implies that any generalisation bound derived for dropout can be regarded as a pessimistic bound for domain-specific, standard data augmentation. A similar argument applies for weight decay, which, as first shown by Bishop (1995), is equivalent to training with noisy examples if the noise amplitude is small and the objective is the sum-of-squares error function. Therefore, some forms of explicit regularisation are at least approximately equivalent to adding random noise to the training examples, which is the simplest form of data augmentation[2]. Thus, it is reasonable to argue that more sophisticated data augmentation can overshadow the benefits provided by explicit regularisation.

In general, we argue that the reason why explicit regularisation may not be necessary is that neural networks are already implicitly regularised by many elements—stochastic gradient descent (SGD), convolutional layers, normalisation and data augmentation, to name a few—that provide a more successful inductive bias (Neyshabur et al., 2014). For instance, it has been shown that linear models optimised with SGD converge to solutions with small norm, without any explicit regularisation (Zhang et al., 2017a). Furthermore, as discussed in Section 3.4, if overparameterised neural networks are able to generalise well, the need for constraining their capacity is questionable. In the rest of the chapter we present an empirical study to contrast data augmentation and explicit regularisation—weight decay and dropout.

4.2 Methods

This section describes the main aspects of the experimental setup for systematically analysing the role of data augmentation in deep neural networks compared to weight decay and dropout.

4.2.1 Data

We performed the experiments on the highly benchmarked data sets ImageNet (Russakovsky et al., 2015) ILSVRC 2012, CIFAR-10 and CIFAR-100 (Krizhevsky & Hinton, 2009). We resized the 1.3 M images from ImageNet into 150×200 pixels, as a compromise between keeping a high resolution and speeding up the training. Both on ImageNet and on CIFAR, the pixel values were mapped into the range $[0, 1]$.

So as to analyse the role of data augmentation, we trained every model with two different augmentation schemes as well as with no data augmentation at all. The two augmentation schemes are the following:

Light augmentation

This scheme is common in the literature, for example (Goodfellow et al., 2013; Springenberg et al., 2014), and performs only horizontal flips and horizontal and vertical translations of 10% of the image size.

Heavier augmentation

This scheme performs a larger range of affine transformations such as scaling, rotations and shear mappings, as well as contrast and brightness adjustment. On ImageNet we additionally per-

[2]Note that the opposite view—domain-specific data augmentation as explicit regularisation—does not apply. In Section 3.3 we discuss about the taxonomies of regularisation, including the difference between data augmentation and data-dependent regularisation

formed random crops of 128×128 pixels. The choice of the allowed transformations is arbitrary and the only criterion was that the objects be still recognisable in general. We deliberately avoided designing a particularly successful scheme. The details of the transformations are presented below and the range of the parameters are specified in Table 4.1 and some visual examples are shown in Figure 4.1.

- Affine transformations:
$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} f_h z_x \cos(\theta) & -z_y \sin(\theta + \phi) & t_x \\ z_x \sin(\theta) & z_y \cos(\theta + \phi) & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$
- Contrast adjustment: $x' = \gamma(x - \overline{x}) + \overline{x}$
- Brightness adjustment: $x' = x + \delta$

Parameter	Description	Range
f_h	Horiz. flip	$1 - 2B(0.5)$
t_x	Horiz. translation	$\mathcal{U}(-0.1, 0.1)$
t_y	Vert. translation	$\mathcal{U}(-0.1, 0.1)$
z_x	Horiz. scale	$\mathcal{U}(0.85, 1.15)$
z_y	Vert. scale	$\mathcal{U}(0.85, 1.15)$
θ	Rotation angle	$\mathcal{U}(-22.5°, 22.5°)$
ϕ	Shear angle	$\mathcal{U}(-0.15, 0.15)$
γ	Contrast	$\mathcal{U}(0.5, 1.5)$
δ	Brightness	$\mathcal{U}(-0.25, 0.25)$

Table 4.1: Description and range of possible values of the parameters used for the heavier augmentation scheme. $B(p)$ denotes a Bernoulli distribution and $\mathcal{U}(a, b)$ a uniform distribution.

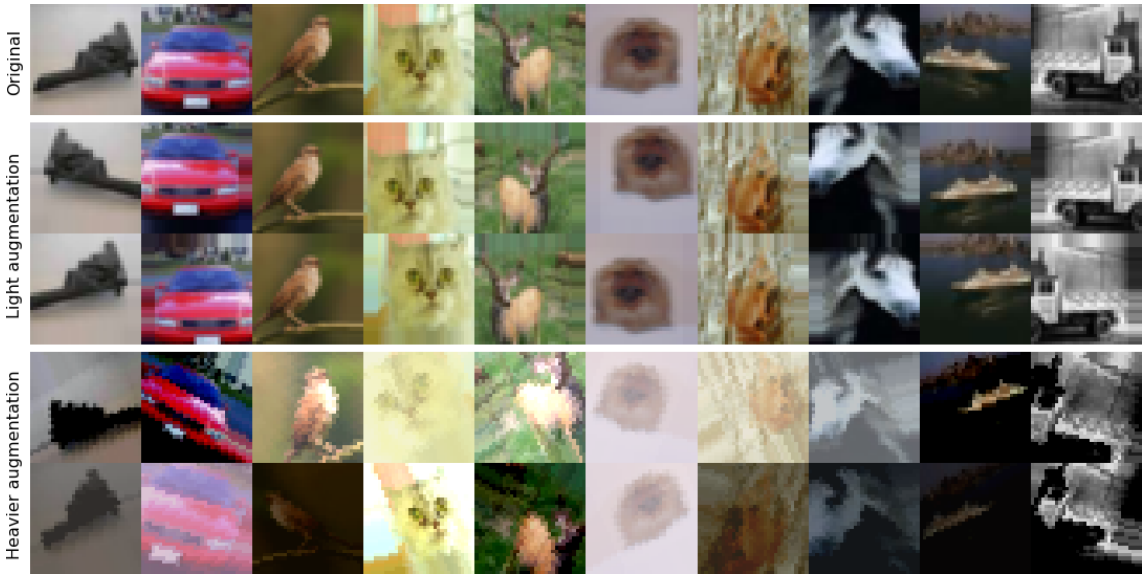


Figure 4.1: Illustration of the most extreme transformations performed by the data augmentation schemes on ten images—one per class—from CIFAR-10.

4.2.2 Network Architectures

We trained three distinct, popular architectures that have achieved successful results in visual object recognition: the all convolutional network, All-CNN (Springenberg et al., 2014); the wide

residual network, WRN (Zagoruyko & Komodakis, 2016); and the densely connected network, DenseNet (Huang et al., 2017). Importantly, we kept the training hyperparameters—learning rate, training epochs, batch size, optimiser, etc.—as in the original papers. Table 4.2 summarises the main features of each network and below we specify further details.

	All-CNN	**WRN**	**DenseNet**
Ref. in original paper	*All-CNN-C*	*WRN-28-10*	*DenseNet-BC*
Main feature	Only conv. layers	Residual connections	Dense connectivity
Number of layers	16 / 12	28	101
Number of parameters	9.4 / 1.3 M	36.5 M	0.8 M
Training hours	35–45 / 2.5	100–145 / 14–15	24–27
CO2e emissions[3](kg)	4.17–5.36 / 0.29	11.91–17.27 / 1.66–1.78	2.86–3.21

Table 4.2: Key aspects of the network architectures. Cells with two values correspond to ImageNet / CIFAR.

All Convolutional Network

All-CNN consists exclusively of convolutional layers with ReLU activation (Glorot et al., 2011), it is relatively shallow and has few parameters. For ImageNet, the network has 16 layers and 9.4 million parameters; for CIFAR, it has 12 layers and about 1.3 million parameters. In our experiments to compare the adaptability of data augmentation and explicit regularisation to changes in the architecture (Section 4.3.3), we also tested a *shallower* version, with 9 layers and 374,000 parameters, and a *deeper* version, with 15 layers and 2.4 million parameters. The four architectures can be described as in Table 4.3, where $K\mathbf{C}D(S)$ is a $D \times D$ convolutional layer with K channels and stride S, followed by batch normalisation and a ReLU non-linearity. *N.Cl.* is the number of classes and Gl.Avg. refers to global average pooling.

ImageNet	96**C**11(2)–96**C**1(1)–96**C**3(2)–256**C**5(1) –256**C**1(1)–256**C**3(2)–384**C**3(1) –384**C**1(1)–384**C**3(2)–1024**C**3(1) –1024**C**1(1)–*N.Cl.*C1(1) –Gl.Avg.–Softmax
CIFAR	2×96**C**3(1)–96**C**3(2)–2×192**C**3(1) –192**C**3(2)–192**C**3(1)–192**C**1(1) –*N.Cl.*C1(1)–Gl.Avg.–Softmax
Shallower	2×96**C**3(1)–96**C**3(2)–192**C**3(1) –192**C**1(1)–*N.Cl.*C1(1)–Gl.Avg.–Softmax
Deeper	2×96**C**3(1)–96**C**3(2)–2×192**C**3(1) –192**C**3(2)–2×192**C**3(1)–192**C**3(2) –192**C**3(1)–192**C**1(1) –*N.Cl.*C1(1)–Gl.Avg.–Softmax

Table 4.3: Specification of the All-CNN architectures.

The CIFAR network is identical to the All-CNN-C architecture in the original paper, except for the introduction of the batch normalisation layers (Ioffe & Szegedy, 2015), which we included because they generally improve performance, but had not been proposed at the time of publication of All-CNN (Springenberg et al., 2014). The ImageNet version also includes batch normalisation layers and a stride of 2 instead of 4 in the first layer to compensate for the reduced input size.

[3]The carbon emissions were computed using the online calculator at green-algorithms.org Lannelongue et al. (2020). The whole set of experiments in this chapter emmited an estimated total of 390.45 CO2e. The details about how the carbon emissions were calculated and about the impact of this study on global warming were made available as supplementary material of the main publication of this chapter .

Importantly, we kept the same training parameters as in the original paper in the cases they were reported. Specifically, the All-CNN networks were trained using stochastic gradient descent, with fixed Nesterov momentum 0.9, learning rate of 0.01 and decay factor of 0.1. The batch size for the experiments on ImageNet was 64 and we trained during 25 epochs decaying the learning rate at epochs 10 and 20. On CIFAR, the batch size was 128, we trained for 350 epochs and decayed the learning rate at epochs 200, 250 and 300. The kernel parameters were initialised according to the Xavier uniform initialisation (Glorot & Bengio, 2010).

Wide Residual Network

WRN is a modification of ResNet (He et al., 2016) that achieves better performance with fewer layers, but more units per layer. Here, we chose for our experiments the WRN-28-10 version (28 layers and about 36.5 M parameters), which was reported to achieve the best results on CIFAR. It has the following architecture:

$$16\mathbf{C}3(1)\text{–}4\times160\mathbf{R}\text{–}4\times320\mathbf{R}\text{–}4\times640\mathbf{R}\text{–BN–ReLU–Avg.(8)–FC–Softmax}$$

where $K\mathbf{R}$ is a residual block with residual function BN–ReLU–$K\mathbf{C}3(1)$–BN–ReLU–$K\mathbf{C}$ 3(1). BN is batch normalisation, Avg.(8) is spatial average pooling of size 8 and FC is a fully connected layer. On ImageNet, the stride of the first convolution is 2. The stride of the first convolution within the residual blocks is 1 except in the first block of the series of 4, where it was set to 2 in order to subsample the feature maps.

Similarly, we kept the training parameters of the original paper: we trained with SGD, with fixed Nesterov momentum 0.9 and learning rate of 0.1. On ImageNet, the learning rate was decayed by 0.2 at epochs 8 and 15 and we trained for a total of 20 epochs with batch size 32. On CIFAR, we trained with a batch size of 128 during 200 epochs and decayed the learning rate at epochs 60, 120 and 160. The kernel parameters were initialised according to the He normal initialisation (He et al., 2015).

DenseNet

The main characteristic of DenseNet (Huang et al., 2017) is that the architecture is arranged into blocks whose layers are connected to all the layers below, forming a dense graph of connections, which permits training very deep architectures with fewer parameters than, for instance, ResNet. Here, we used a network with bottleneck compression rate $\theta = 0.5$ (DenseNet-BC), growth rate $k = 12$ and 16 layers in each of the three blocks. The model has nearly 0.8 million parameters. The specific architecture can be described as follows:

$$2\times k\mathbf{C}3(1)\text{–DB(16)–TB–DB(16)–TB–DB(16)–BN–Gl.Avg.–FC–Softmax}$$

where DB(c) is a dense block, that is a concatenation of c convolutional blocks. Each convolutional block is a set of layers whose output is concatenated with the input to form the input of the next convolutional block. A convolutional block with bottleneck structure has the following layers:

$$\text{BN–ReLU–}4\times k\mathbf{C}1(1)\text{–BN–ReLU–}k\mathbf{C}3(1)\text{–Concat.}$$

TB is a transition block, which downsamples the size of the feature maps, formed by the following layers:

$$\text{BN–ReLU–}k\mathbf{C}1(1)\text{–Avg.(2).}$$

Like with All-CNN and WRN, we kept the training hyperparameters of the original paper. On the CIFAR data sets, we trained with SGD, with fixed Nesterov momentum 0.9 and learning rate of 0.1, decayed by 0.1 on epochs 150 and 200 and training for a total of 300 epochs. The batch size was 64 and the parameters were initialised with He initialisation.

4.2.3 Train and Test

Every architecture was trained on each data set both with explicit regularisation—weight decay and dropout as specified in the original papers—and without. Furthermore, we trained each model with the three data augmentation schemes: no augmentation, light and heavier. Figure 4.2 shows a summary of this experimental setup. The performance of the models was computed on the held out test tests. As in previous works (Krizhevsky et al., 2012; Simonyan & Zisserman, 2014), we averaged the softmax posteriors over 10 random *light* augmentations, since slightly better results are obtained. Then we computed the classification accuracy for the models trained on CIFAR and the top-5 accuracy for the ImageNet models.

All the experiments were performed on Keras (Chollet et al., 2015) on top of TensorFlow (Abadi et al., 2015), with a single GPU NVIDIA GeForce GTX 1080 Ti.

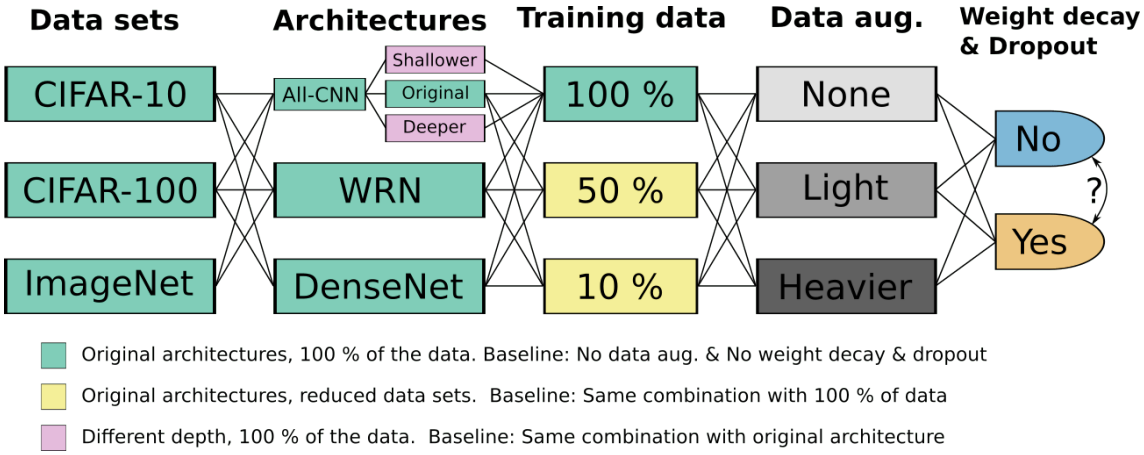


Figure 4.2: Visual summary of the experimental setup. The figure represents the factors of variation in our experiments: data sets, architectures, amount of training data, data augmentation scheme and inclusion of explicit regularization. Comparisons within a factor of variation are most relevant on the factors on the right, like the performance of the models train with and without explicit regularization.

4.2.4 Carbon footprint of the computational experiments

Training artificial neural networks effectively on large, non-trivial data sets consumes a considerable amount of energy (Strubell et al., 2019). Crucially, the amount of compute of the largest models has been increasing exponentially during the last decade (Amodei & Hernandez, 2018). Therefore, the contribution of deep learning research to global warming and climate change should not be neglected (Schwartz et al., 2019; Lacoste et al., 2019; Lannelongue et al., 2020). As the experimental design of this chapter required training multiple neural network models, we wanted to be both aware and transparent about the environmental impact of our experimental study and in this section we will report calculations of the estimated carbon emissions associated with training our models for this chapter and the details of how these are computed.

In Table 4.2 of Section 4.2.2 we report the estimated carbon emissions associated with training each architecture on each data set, given the specific characteristics of our computing hardware. These estimations rely on the online calculator available at green-algorithms.org, developed by Lannelongue et al. (2020). In order to estimate the carbon emissions of each model, we took into consideration the following information: all the models were trained in a local desktop computer, located in Germany, with a single graphic processing unit (GPU), model GTX 1080 Ti, with 11 GB of memory. We assumed full usage of the 11 GB of memory and of the processing core for all models—a conservative estimation.

Table 4.4: Summary of estimated carbon emissions associated to training the models for our experimental setup

Network	Data set	Depth	% Data	N. models	Total h.	Total CO2e
All-CNN	CIFAR 2.5 h 0.29 CO2e	original	100 %	36	90	10.73
			50 %	24	30	3.58
			10 %	24	6	0.72
		shallower	100 %	12	25.2	3.0
		deeper	100 %	12	36	4.29
	ImageNet 35–45 h 4.17–5.36 CO2e	original	100 %	6	270	32.18
			50 %	6	135	16.09
			10 %	6	27	3.22
WRN	CIFAR 14–15 h 1.66–1.78 CO2e	original	100 %	36	540	64.35
			50 %	12	90	10.73
			10 %	12	18	2.15
	ImageNet 100–145 h 11.91–17.27 CO2e	original	100 %	6	870	103.68
			50 %	6	435	51.84
			10 %	6	87	10.37
DenseNet	CIFAR 24–27 h 2.86–3.21 CO2e	original	100 %	12	324	38.61
			50 %	6	81	9.65
			10 %	6	16.2	1.93
	Total			**228**	**3080.4**	**367.12**

As reported in Table 4.4, the complete set of experiments reported in this chapter needed a total of 3,276 GPU hours (136.5 days), which correspond to actual real time, since we had access to a single GPU. With our hardware, this corresponds, according to Lannelongue et al. (2020), to an estimate of 832.53 kWh or 390.45 carbon dioxide equivalent (CO2e). Carbon dioxide equivalent represents the equivalent CO2 that would have the same global warming impact than a mixture of gases. By way of comparison, 390.45 CO2e corresponds to 34.25 tree-years—the time taken by a mature tree to absorb the CO2—, 69 % of a flight New York City–San Francisco or 2,231 km in a passenger car.

4.3 Results

Here we present the results of the empirical study. In the first set of experiments (Section 4.3.1) we trained the architectures as in the original papers with the full data sets. A relevant characteristic of explicit regularisation methods is that they typically require the specification of hyperparameters. These are usually fine-tuned by the authors of research papers to achieve higher performance, as demanded by the dynamics of the scientific publication environment in the machine learning community. However, the sensitivity of the results towards these hyperparameters is often not made available. In order to gain insight on the role of explicit regularisation and data augmentation in more real world cases, where the hyperparameters have not been highly optimised, we varied the amount of training data (Section 4.3.2) and the depth of the architectures (Section 4.3.3), while keeping all other hyperparameters untouched.

The objective of the study is to contrast the performance gained by training the models with both explicit regularisation and data augmentation, which is the common practice in the literature (Tan & Le, 2019; Huang et al., 2017; Zagoruyko & Komodakis, 2016; Springenberg et al., 2014), against training with only data augmentation. Hence, the presentation of the results in the figures aims at facilitating this comparison. In the performance plots, we represent the relative performance gain of each model with respect to the relevant baseline, which we specify at each section. We plot the results in pairs: the squared blue dots on the top, blue-shaded area correspond to the models trained with only data augmentation and the round orange dots on

the bottom, orange-shaded area to the models trained with both data augmentation and explicit regularization. Additionally, the results of training with different levels of data augmentation are represented with dots in three lightness and saturation shades and we connected with dotted lines the models trained with the same level of augmentation.

In order to assess the statistical significance of the differences between models trained with and without explicit regularisation, we carried out percentile bootstrap analyses (Efron, 1992), that is simulations based on sampling with replacement. We followed the guidelines by Rousselet et al. (2019). In all cases, the values of the distribution correspond to the difference between the performance—with respect to the baseline—of the models trained without explicit regularisation minus the performance of the models trained with explicit regularisation—the pairs of dots connected by a dotted line. We then compared the distribution of this difference in the bootstrap samples and with respect to the null hypothesis, that is no difference ($H_0 = 0$). For each experiment we sampled all possible bootstrap samples with replacement or a maximum of one million.

4.3.1 Original architectures

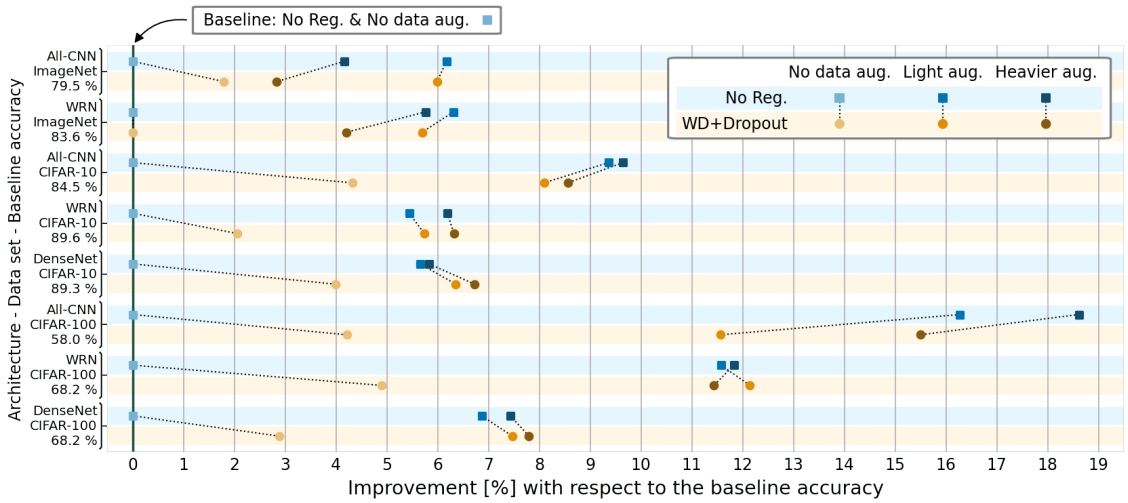


Figure 4.3: Relative improvement of adding data augmentation and explicit regularization to the baseline models, $(accuracy - baseline)/accuracy * 100$. The baseline accuracy is shown on the left. The results suggest that data augmentation alone (in blue) can achieve even better performance than the models trained with both weight decay and dropout (in orange).

First, we contrast the regularisation effect of data augmentation and weight decay and dropout on the original networks trained with the complete data sets, and show the results in Figure 4.3. As a baseline, we consider the "bare bone" models, that is the model trained with neither explicit regularisation nor data augmentation. We report the accuracy of the baseline on the left axis of the plot in Figure 4.3. To assess the relevant comparisons, we show the relative improvement in test performance achieved by adding each technique or combination of techniques to the baseline model. Table 4.5 shows the mean and standard deviation of each combination on the architecture and data set and Figure 4.4 the results of the bootstrap analysis, which considers the differences of all pairs—squared blue dots minus round orange dots, connected with dotted lines[4].

The first conclusion from Figures 4.3 and 4.4 as well as Table 4.5 is that training with data augmentation alone (blue dots on the top, blue-shaded areas) is better than training with both augmentation and explicit regularisation (in orange). This is the case in more than half of the

[4]The relative performance of WRN on ImageNet trained with weight decay and dropout with respect to the baseline is negative (-6.22 %) and is neither depicted in Figure 4.3 nor taken into consideration to compute the average improvements in Table 4.5 and the bootstrap analysis in Figure 4.4.

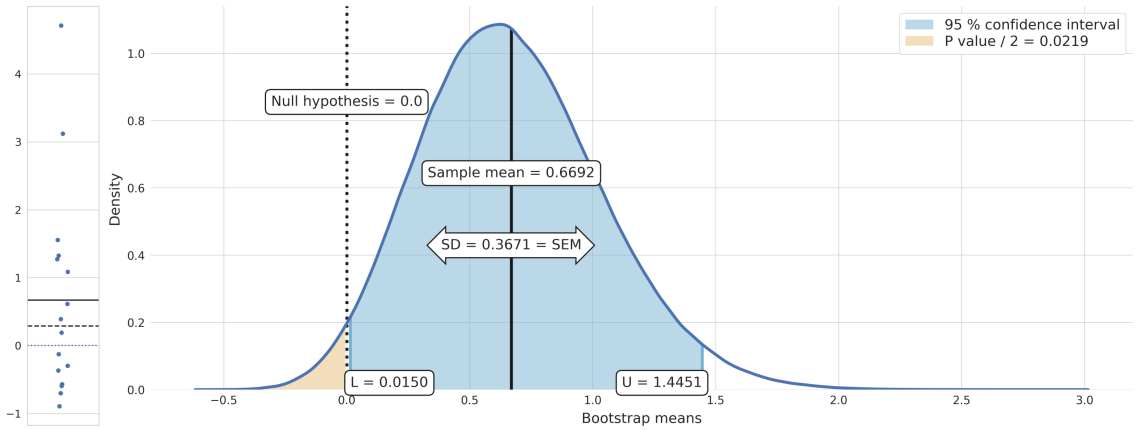


Figure 4.4: Bootstrap analysis to assess the difference in performance gain provided by training without and with weight decay and dropout, on the original architectures and using the full data sets. On the left of the figure we plot the bootstrap values—differences—with the mean and median as a solid and dashed line, respectively. The main figure shows the distribution of the mean of the bootstrap samples, the standard error of the sample mean, the 95 % confidence intervals and the P value with respect to the null hypothesis ($H_0 = 0$).

cases (9/16) and the bootstrap analysis reveals that the difference is positive with 95 % confidence and P value $= 0.022$. On average, adding data augmentation to the baseline model improved the accuracy on 8.57 %, and adding both augmentation and explicit regularisation on 7.90 % respectively.

At first glance, one may think that this is not remarkable, since the differences are small and data augmentation alone is not better in 100 % of the cases. However, this result is surprising and remarkable for the following reason: note that the studied architectures achieved state-of-the-art results at the moment of their publication and the models included all light augmentation, weight decay and dropout, whose parameters were presumably finely tuned to optimise the accuracy. The replication of these results corresponds to the mid-orange dots in Figure 4.3. Here, we have show that simply removing weight decay and dropout—while keeping all other hyperparameters intact, see Section 4.2.2—improves the *then state-of-the-art* accuracy in 4 of the 8 studied cases. Why did not the authors trained without explicit regularisation and obtain better results?

Second, it can also be observed that the regularisation effect of weight decay and dropout, an average improvement of 3.02 % with respect to the baseline, is much smaller than that of data augmentation: simply applying light augmentation increased the accuracy in 8.46 % on average. Although the heavier augmentation scheme was deliberately not designed to optimise the performance, in both CIFAR-10 and CIFAR-100 it improved the test performance with respect to the light augmentation scheme. This was not the case on ImageNet, probably due to the larger complexity of the data set.

Further, it can be observed that the results are in general more consistent in the models trained without explicit regularisation. Finally, an additional advantage of training without explicit regularisation is that the learning dynamics (Figure 4.5) is much faster and predictive of the final performance. Typically, regularisers such as weight decay and dropout effectively prevent the

	No explicit reg.	Weight decay + dropout
None	*baseline*	3.02 (1.65)
Light	8.46 (3.80)	7.88 (2.60)
Heavier	8.68 (4.69)	7.92 (4.03)

Table 4.5: Average accuracy improvement over the baseline model of each combination of data augmentation level and presence of weight decay and dropout.

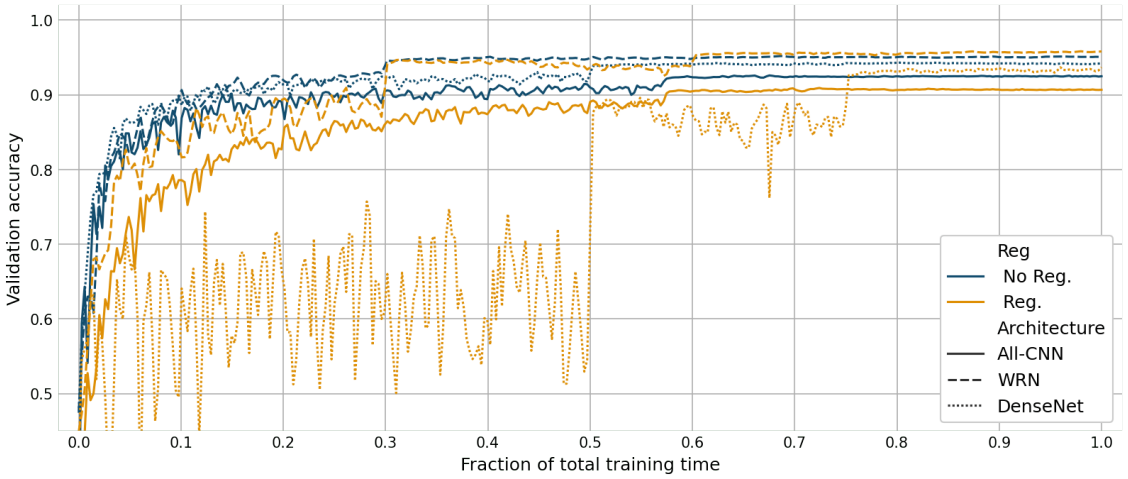


Figure 4.5: Dynamics of the validation accuracy during training of All-CNN, WRN and DenseNet, trained on CIFAR-10 with heavier data augmentation, contrasting the models trained with explicit regularization (orange lines) and the models trained with only data augmentation (in blue). The regularized models heavily rely on the learning rate decay to obtain the boost of performance, while the models trained without explicit regularization quickly approach the final performance.

model from fitting the training data during the first epochs and heavily rely on the learning rate decay to obtain the boost that yields the final performance. On the contrary, models trained with only data augmentation reach very high validation performance after a few epochs. This effect is particularly acute on DenseNet, which performs heavier weight decay.

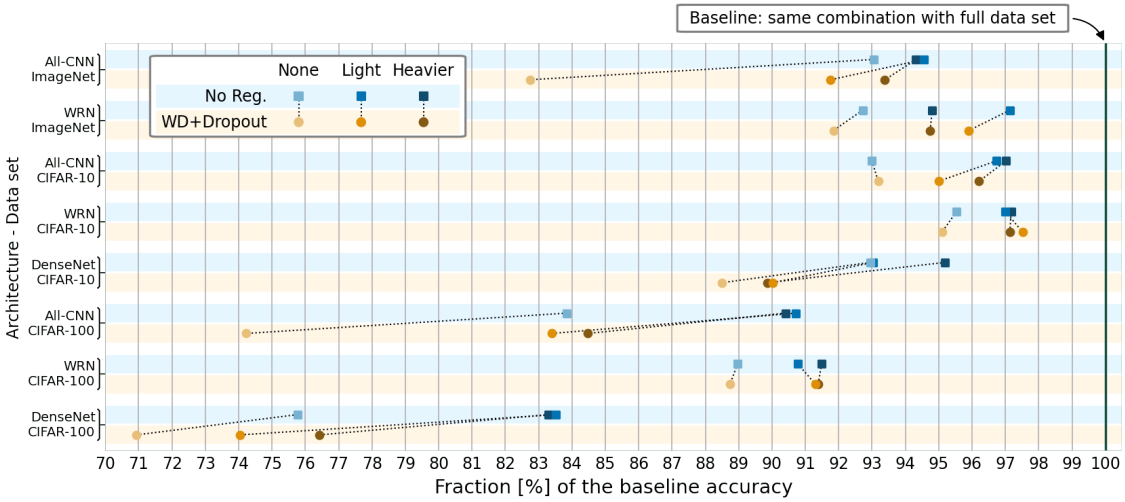
In sum, it seems the performance gain provided by weight decay and dropout can be achieved and often improved by data augmentation alone. Besides, the models trained without explicit regularisation presented additional advantages, which we will further discuss in Section 4.4.

4.3.2 When the available training data changes

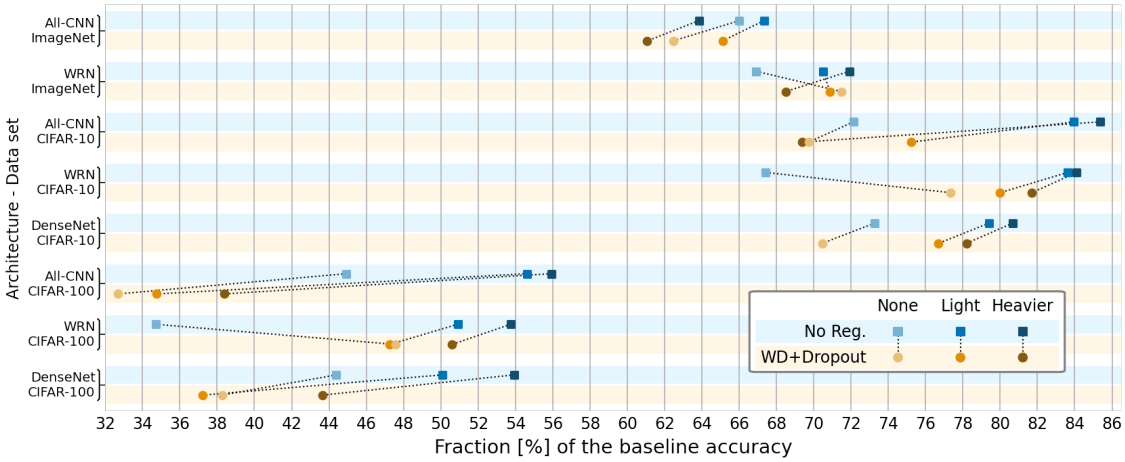
We argue that one of the main drawbacks of explicit regularisation techniques is their poor adaptability to changes in the conditions with which the hyperparameters were tuned. To test this hypothesis and contrast it with the adaptability of data augmentation, we extended the analysis by training the same networks with fewer examples. All models were trained with the same random subset of data and evaluated in the same test set as the previous experiments. In order to better visualise how well each technique resists the reduction of training data, in Figure 4.6 we show the fraction of baseline accuracy achieved by each model when trained with 50 % and 10 % of the available data. In this case, the baseline is thus each corresponding model trained with the complete data set. Table 4.6 summarises the mean and standard deviation of each combination and Figure 4.7 shows the result of the bootstrap analysis.

One of the main conclusions of this set of experiments is that if no data augmentation is applied, explicit regularisation hardly resists the reduction of training data by itself. On average, with 50 % of the available data, these models only achieve 83.20 % of the original accuracy (Table 4.6), which, remarkably, is even worse than the models trained without any explicit regularisation (88.11 %). On 10 % of the data, the average fraction is the same (58.75 and 58.72 %, respectively). This implies that training with explicit regularisation is even detrimental for the performance.

When combined with data augmentation, the models trained with explicit regularisation (orange dots) also perform worse (88.78 and 61.16 % with 50 and 10 % of the data, respectively), than the models without explicit regularisation (blue dots, 91.64 and 68.12 % on average). Note that the difference becomes larger as the amount of available data decreases. Even more decisive are the



(a) 50 % of the available training data



(b) 10 % of the available training data

Figure 4.6: Fraction of the baseline performance when the amount of available training data is reduced, $accuracy/baseline * 100$. The models trained wit explicit regularisation present a significant drop in performance as compared to the models trained with only data augmentation. The differences become larger as the amount of training data decreases.

results of the bootstrap analysis (Figure 4.7): the mean difference of the fraction of the performance achieved by the models trained without and with explicit regularisation is 2.78 and 6.96, with 50 and 10 % of the training data, respectively; the confidence intervals are well above the null hypothesis and the P values are exactly 0.

Importantly, it seems that the combination of explicit regularisation and data augmentation is only slightly better than training without data augmentation. Two reasons may explain this: first, the original regularisation hyperparameters seem to adapt poorly to the new conditions. The hyperparameters were specifically tuned for the original setup and they would require re-tuning to obtain comparable results. Second, since explicit regularisation reduces the representational capacity, this might prevent the models from taking advantage of the augmented data.

In contrast, the models trained without explicit regularisation but with data augmentation more naturally adapt to the reduced availability of data. With 50 % of the data, these models achieve about 91.5 % of the performance with respect to training with the complete data sets. With only 10 % of the data, they achieve nearly 70 % of the baseline performance, on average. This highlights the suitability of data augmentation to serve, to a great extent, as true, useful data

	50 % of the training data	
	No explicit reg.	Weight decay + dropout
None	88.11 (6.27)	83.20 (9.83)
Light	91.47 (4.31)	88.27 (7.39)
Heavier	91.82 (4.63)	89.28 (6.63)
	10 % of the training data	
	No explicit reg.	Weight decay + dropout
None	58.72 (14.93)	58.75 (16.92)
Light	67.55 (14.27)	60.89 (18.39)
Heavier	68.69 (13.61)	61.43 (15.90)

Table 4.6: Average fraction of the original accuracy of each corresponding combination of data augmentation level and presence of weight decay and dropout.

(Vinyals et al., 2016).

4.3.3 When the architecture changes

Finally, in the same spirit, we tested the adaptability of data augmentation and explicit regularisation to changes in the depth of the All-CNN architecture, by training shallower (9 layers) and deeper (15 layers) versions of the architecture. We show the fraction of the performance with respect to the original architecture in Figure 4.8 and the bootstrap analysis in Figure 4.9.

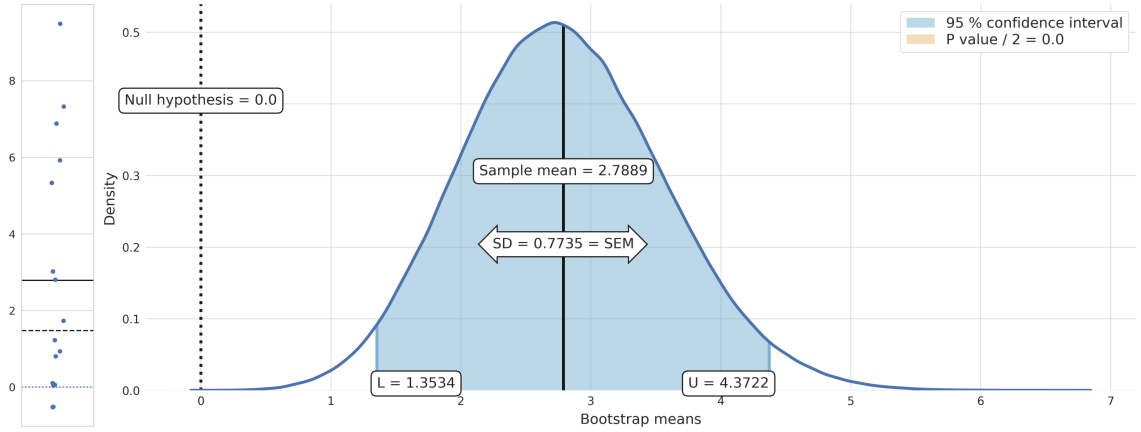
A noticeable result from these experiments is that all the models trained with weight decay and dropout (round orange dots) suffered a dramatic drop in performance when the architecture changed, regardless of whether deeper or shallower and of the amount of data augmentation. As a matter of fact, the models trained without explicit regularisation performed on average 11.23 % ($SD = 3.06$) better. As in the case of reduced training data, this may be explained by the poor adaptability of the regularisation hyperparameters, which strongly depend on the architecture.

This highly contrasts with the performance of the models trained without explicit regularisation (top, squared blue dots). With a deeper architecture, these models achieve slightly better performance, effectively exploiting the increased capacity. With a shallower architecture, they achieve only slightly worse performance[5]. Thus, these models seem to more naturally adapt to the new architecture and data augmentation becomes beneficial.

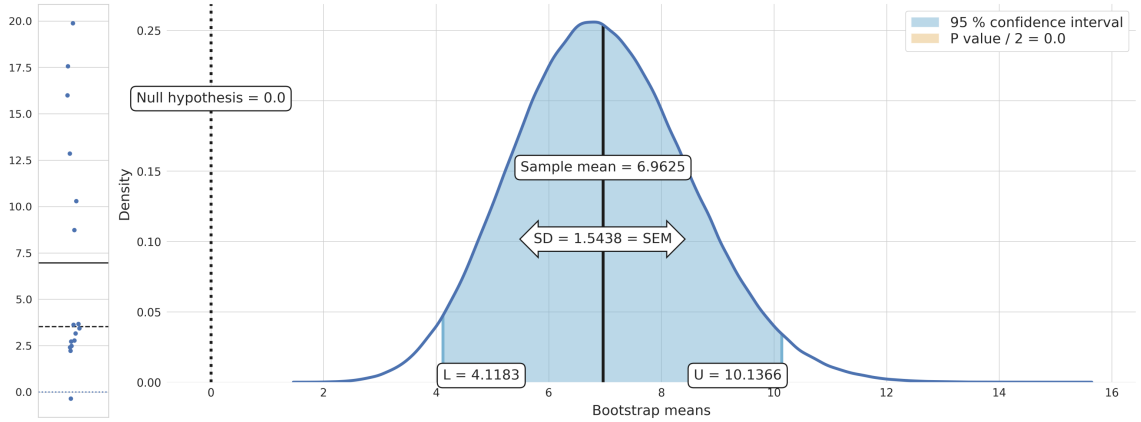
It is worth commenting on the particular case of the CIFAR-100 benchmark, where the difference between the models with and without explicit regularisation is even more pronounced, in general. It is common practice in object recognition papers to tune the parameters for CIFAR-10 and then test the performance on CIFAR-100 with the same hyperparameters. Therefore, these are typically less suitable for CIFAR-100. We believe this is the reason why the benefits of data augmentation seem even more pronounced on CIFAR-100 in our experiments.

In sum, these results highlight another crucial advantage of data augmentation: the effectiveness of its hyperparameters, that is the type of image transformations, depend mostly on the type of data, rather than on the particular architecture or amount of available training data, unlike explicit regularisation hyperparameters. Therefore, removing explicit regularisation and training with data augmentation increases the flexibility of the models.

[5]Note that the shallower models trained with neither explicit regularisation nor data augmentation achieve even better accuracy than their counterpart with the original architecture, probably due to the reduction of overfitting provided by the reduced capacity.



(a) 50 % of the available training data



(b) 10 % of the available training data

Figure 4.7: Bootstrap analysis analogous to the one detailed in Section 4.3.1 and Figure 4.4, to analyse the statistical significance of the performance difference of models trained with only 10 and 50 % of the data.

4.4 Discussion

In this section we summarise our findings and discuss their relevance. In particular, we challenge the need for weight decay and dropout to train artificial neural networks, and propose to rethink data augmentation as a *first class* technique instead of a *cheating* method.

As an empirical analysis, one caveat of our work is the limited number of experiments (over 200 models trained). In order to increase the generality of our conclusions, we chose three significantly distinct network architectures and three data sets. Importantly, we also took a conservative approach in our experimentation: all the hyperparameters were kept as in the original models, which included both weight decay and dropout, as well as light augmentation. This setup is clearly suboptimal for models trained without explicit regularisation. Besides, the heavier data augmentation scheme was deliberately not optimised to improve the performance as it was not the scope of this work to propose a specific data augmentation technique. We leave for future work to explore data augmentation schemes that can more successfully be exploited by any deep model. Finally, in order to strengthen the conclusions from the empirical analysis, we have also discussed some theoretical insights in Section 4.1, concluding that the generalisation gain provided by weight decay can be seen as a lower bound of what can be achieved by domain-specific data augmentation. We also hope that this work inspires researchers in other application domains, such as natural language processing, to further contrast data augmentation and explicit regularisation.

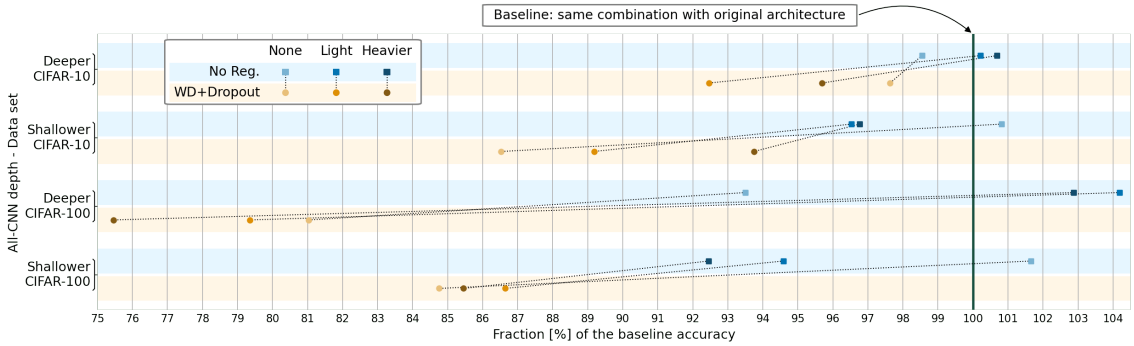


Figure 4.8: Fraction of the original performance when the depth of the All-CNN architecture is increased or reduced in 3 layers. In the explicitly regularised models, the change of architecture implies a dramatic drop in the performance, while the models trained without explicit regularisation present only slight variations with respect to the original architecture.

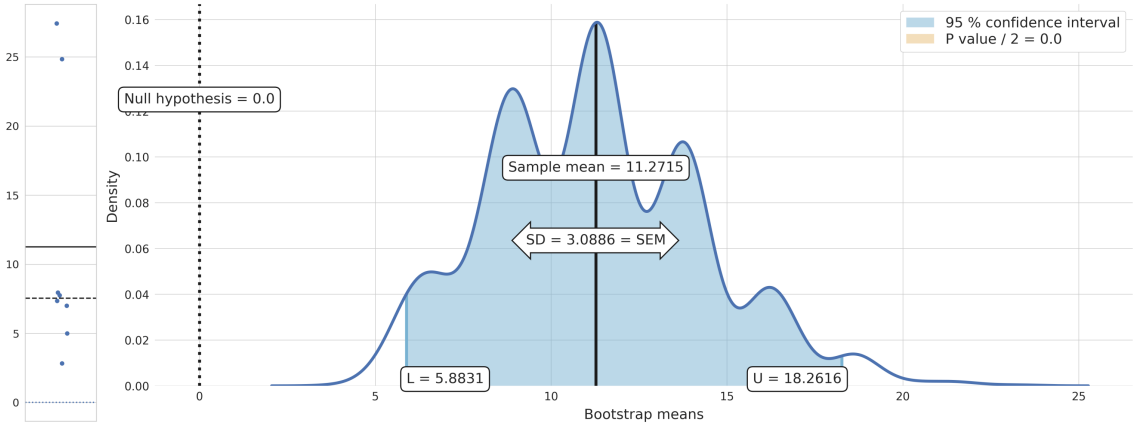


Figure 4.9: Bootstrap analysis analogous to the one detailed in Section 4.3.1 and Figure 4.4, to analyse the statistical significance of the performance difference of All-CNN trained with 3 more and 3 fewer layers.

4.4.1 Do deep nets really need weight decay and dropout?

In Section 4.3 we have presented the results of a systematic empirical analysis of the role of weight decay, dropout and data augmentation in deep convolutional neural networks for object recognition. Our results have shown that explicit regularisation is not only unnecessary (Zhang et al., 2017a), but also that its performance gain can be achieved by data augmentation alone: in most cases, training with data augmentation only was better than training with both data augmentation and explicit regularisation. In the few cases where that was not the case, the difference was very small. Moreover, unlike data augmentation, models trained with weight decay and dropout exhibited poor adaptability to changes in the architecture and the amount of training data. Why do researchers and practitioners keep training their neural networks with weight decay and dropout? Do deep nets really need weight decay and dropout?

The relevance of these findings lies in the fact that weight decay and dropout are almost ubiquitously present in convolutional neural networks (Huang et al., 2017; Zagoruyko & Komodakis, 2016; Springenberg et al., 2014), including recent, state of the art models (Tan & Le, 2019). Certainly, it has been shown in multiple research papers that weight decay and dropout can boost the performance of neural networks, and we here do not challenge the usefulness of weight decay and dropout, but the convenience to use it, given the associated cost and risk, and available alternatives.

First, not only add weight decay and dropout extra computations during training, but also they typically require training the models several times with different hyperparameters: the coefficient of the penalty for weight decay; for dropout, the location of the dropout mask and the amount of units to drop. These hyperparameters are arguably very sensitive to changes in elements of the learning process. Here we have studied changes in the amount of training data (Section 4.3.2) and the depth of the architecture (Section 4.3.3). Consider, for instance, our results in Section 4.3.2: All-CNN trained on CIFAR-10 with weight decay, dropout and light augmentation reaches about 92 % accuracy. If we were in the development process and were unsure about what architecture to use, we could simply try our network with three more layers and we would obtain about 85 % accuracy. We could also try an architecture with three fewer layers and obtain about 82 % accuracy. This may lead us to conclude that the first architecture has the right number of layers, because adding or removing layers drastically reduces the performance; or perhaps that by adding or removing layers there is some negative interaction between the layers sizes, or any other of the many hypothesis we could think of.

Consider now what happens if we train without weight decay and dropout: All-CNN trained on CIFAR-10 with light augmentation—but without weight decay and dropout—obtains about 93.3 % accuracy. This is slightly better than the explicitly regularised model, but we will ignore this now. If we train this model with three more layers, we obtain 93.4 % accuracy, that is the same or slightly better—as opposed to the drop of 7 points we have seen before. If we train with three fewer layers, we obtain about 90 % accuracy, a drop of 3 points—as opposed to a drop of 10 points. In this case, we would not conclude that adding or removing layers creates negative interactions. Note that the only difference between these two cases is that the first models are trained with weight decay and dropout. Therefore, it may be reasonable to only include explicit regularisation in the final version of a model, in order to potentially obtain a slight boost in performance prior to publication or production—provided the hyperparameters are adequately fine-tuned—but keeping weight decay and dropout as intrinsic part of our models can certainly lead us astray.

Finally, we can draw some connections between the results from this chapter and the insights from the previous chapter. In Chapter 3 we discussed that the role of explicit regularisation techniques, such as weight decay and dropout, is to reduce the representational capacity of the models. This, according to statistical learning theory, can reduce overfitting and in turn improve generalisation. However, artificial neural networks have usually orders of magnitude more parameters than training examples, and they still generalise well. While this phenomenon is still not well understood, one working hypothesis is that overparameterisation does not cause negative overfitting, but rather smooth fitting that can be suitable for accurate interpolation (Belkin et al., 2019; Hasson et al., 2020). If overparameterisation is not a problem for artificial neural networks, is it then necessary to constrain the representational capacity through explicit regularisation? Is it reasonable to train very large models, that require a lot of memory and computation—and negatively impact the environment—and at the same time constrain their capacity?

We also hypothesise that a reason why artificial neural networks generalise well in many tasks is due to the fact that the models include many sources of implicit regularisation or, in other words, inductive biases. For example, it is known that stochastic gradient descent naturally converges to solutions with small norm (Zhang et al., 2017a; Neyshabur et al., 2014), batch normalisation also contributes to better generalisation, convolutional layers are particularly efficient to process image data—and not only—to name a few examples. In our case, we argue that data augmentation has the potential to encode very powerful inductive biases that improve generalisation. We conclude that in the presence of many other sources of implicit regularisation and more effective inductive biases, weight decay and dropout may not be necessary to train large deep artificial neural networks[6].

[6]Previous work has suggested interesting connections between weight decay and other types of regularisation and improved adversarial robustness (Galloway et al., 2018; Jakubovitz & Giryes, 2018). An interesting avenue for future work is studying whether this effects are also provided by data augmentation

4.4.2 Rethinking Data Augmentation

Data augmentation is often regarded by authors of machine learning papers as *cheating*, suggesting it should not be used in order to test the potential of newly proposed methods (Goodfellow et al., 2013; Graham, 2014; Larsson et al., 2016). In contrast, weight decay and dropout are considered intrinsic elements of the algorithms (Tan & Le, 2019). In view of our results, we propose to rethink data augmentation and switch roles with explicit regularisation: good models should effectively exploit data augmentation and explicit regularisation should only be applied, if at all, once all other elements are fixed. This approach improves the performance and saves computational resources.

In this regard, it is worth highlighting some advantages of data augmentation: Not only does it not reduce the representational capacity, unlike explicit regularisation, but also, since the transformations reflect plausible variations of the real objects, it increases the robustness of the model Novak et al. (2018); Rusak et al. (2020). Interestingly, in Chapter 5 we will also show that models trained with heavier data augmentation learn representations more aligned with the inferior temporal (IT) cortex, highlighting its connection with visual perception and biological vision. Deep nets are especially well suited for data augmentation because they do not rely on pre-computed features. Moreover, unlike explicit regularisation, it can be performed on the CPU, in parallel to the gradient updates. Finally, from Sections 4.3.2 and 4.3.3 we concluded that data augmentation naturally adapts to architectures of different depth and amounts of available training data, without the need for specific fine-tuning of hyperparameters.

A commonly cited disadvantage of data augmentation is that it depends on expert knowledge and it cannot be applied to all domains (DeVries & Taylor, 2017b). However, we argue instead that expert and domain knowledge should not be disregarded but exploited. Expert and domain knowledge are, in fact, useful inductive biases. A remarkable advantage of data augmentation is that a single augmentation scheme can be designed for a broad family of data—for example, natural images, using our knowledge about visual perception—and effectively applied to a broad set of tasks—object recognition, segmentation, localisation, etc. We hope that these insights encourage more research on data augmentation and, in general, highlight the importance of using the available data more effectively. In the following chapters, we explore additional properties of models trained with data augmentation (Chapter 5) and how it can be used as part of the objective function to learn representations more aligned with the properties of the primate visual cortex (Chapter 6).

Bibliography

Abadi, M. et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

Abu-Mostafa, Y. S. Learning from hints in neural networks. *Journal of Complexity*, 1990.

Amodei, D. and Hernandez, D. AI and compute. Accessed: 2020-09-28, 2018. URL `https://openai.com/blog/ai-and-compute/`.

Antoniou, A., Storkey, A., and Edwards, H. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.

Bartlett, P. L. and Mendelson, S. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research (JMLR)*, 2002.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences (PNAS)*, 2019.

Bishop, C. M. Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, 1995.

Bouthillier, X., Konda, K., Vincent, P., and Memisevic, R. Dropout as data augmentation. *arXiv preprint arXiv:1506.08700*, 2015.

Chen, S., Dobriban, E., and Lee, J. H. Invariance reduces variance: Understanding data augmentation in deep learning and beyond. *arXiv preprint arXiv:1907.10905*, 2019.

Chollet, F. et al. Keras. `https://github.com/fchollet/keras`, 2015.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. AutoAugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.

DeVries, T. and Taylor, G. W. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017a.

DeVries, T. and Taylor, G. W. Dataset augmentation in feature space. In *International Conference on Learning Representations (ICLR), arXiv:1702.05538*, 2017b.

Efron, B. Bootstrap methods: another look at the jackknife. In *Breakthroughs in Statistics*, pp. 569–593. Springer, 1992.

Galloway, A., Tanay, T., and Taylor, G. W. Adversarial training versus weight decay. *arXiv preprint arXiv:1804.03308*, 2018.

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

Glorot, X., Bordes, A., and Bengio, Y. Deep sparse rectifier neural networks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep learning*. MIT press, 2016.

Goodfellow, I. J., Warde-Farley, D., Mirza, M., Courville, A. C., and Bengio, Y. Maxout networks. In *International Conference on Machine Learning (ICML)*, 2013.

Graham, B. Fractional max-pooling. *arXiv preprint arXiv:1412.6071*, 2014.

Hanson, S. J. and Pratt, L. Y. Comparing biases for minimal network construction with back-propagation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1989.

Hasson, U., Nastase, S. A., and Goldstein, A. Direct fit to nature: An evolutionary perspective on biological and artificial neural networks. *Neuron*, 2020.

Hauberg, S., Freifeld, O., Larsen, A. B. L., Fisher, J., and Hansen, L. Dreaming more data: Class-dependent distributions over diffeomorphisms for learned data augmentation. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2016.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Helmbold, D. P. and Long, P. M. Surprising properties of dropout in deep networks. *Journal of Machine Learning Research (JMLR)*, 2017.

Hernández-García, A. and König, P. Data augmentation instead of explicit regularization, 2018. URL `https://openreview.net/forum?id=ByJWeR1AW`.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, 2015.

Jakubovitz, D. and Giryes, R. Improving DNN robustness to adversarial attacks using Jacobian regularization. In *European Conference on Computer Vision (ECCV)*, 2018.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. *Technical report, University of Toronto*, 2009.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012.

Lacoste, A., Luccioni, A., Schmidt, V., and Dandres, T. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019.

Lannelongue, L., Grealey, J., and Inouye, M. Green algorithms: Quantifying the carbon emissions of computation. *arXiv preprint arXiv:2007.07610*, 2020.

Larsson, G., Maire, M., and Shakhnarovich, G. FractalNet: Ultra-deep neural networks without residuals. *arXiv preprint arXiv:1605.07648*, 2016.

Lemley, J., Bazrafkan, S., and Corcoran, P. Smart augmentation-learning an optimal data augmentation strategy. *IEEE Access*, 2017.

Lyle, C., van der Wilk, M., Kwiatkowska, M., Gal, Y., and Bloem-Reddy, B. On the benefits of invariance in neural networks. *arXiv preprint arXiv:2005.00178*, 2020.

Mou, W., Zhou, Y., Gao, J., and Wang, L. Dropout training, data-dependent regularization, and generalization bounds. In *International Conference on Machine Learning (ICML)*, 2018.

Neyshabur, B., Tomioka, R., and Srebro, N. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *International Conference on Learning Representations (ICLR), arXiv:1412.6614*, 2014.

Novak, R., Bahri, Y., Abolafia, D. A., Pennington, J., and Sohl-Dickstein, J. Sensitivity and generalization in neural networks: an empirical study. In *International Conference on Learning Representations (ICLR), arXiv:1802.08760*, 2018.

Perez, L. and Wang, J. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017.

Rajput, S., Feng, Z., Charles, Z., Loh, P.-L., and Papailiopoulos, D. Does data augmentation lead to positive margin? In *International Conference on Machine Learning (ICML)*, 2019.

Ratner, A. J., Ehrenberg, H. R., Hussain, Z., Dunnmon, J., and Ré, C. Learning to compose domain-specific transformations for data augmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Rousselet, G., Pernet, C., and Wilcox, R. R. A practical introduction to the bootstrap: a versatile method to make inferences by using data-driven simulations. *PsyArXiv preprint PsyArXiv:h8ft7*, 2019.

Rusak, E., Schott, L., Zimmermann, R., Bitterwolf, J., Bringmann, O., Bethge, M., and Brendel, W. Increasing the robustness of DNNs against image corruptions by playing the game of noise. *arXiv preprint arXiv:2001.06057*, 2020.

Russakovsky, O. et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 2015.

Schwartz, R., Dodge, J., Smith, N. A., and Etzioni, O. Green AI. *arXiv preprint arXiv:1907.10597*, 2019.

Shorten, C. and Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *Journal of Big Data*, 2019.

Simard, P., Victorri, B., LeCun, Y., and Denker, J. Tangent prop-a formalism for specifying selected invariances in an adaptive network. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1992.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. Striving for simplicity: The all convolutional net. In *International Conference on Learning Representations (ICLR), arXiv:1412.6806*, 2014.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research (JMLR)*, 2014.

Strubell, E., Ganesh, A., and McCallum, A. Energy and policy considerations for deep learning in NLP. *arXiv preprint arXiv:1906.02243*, 2019.

Tan, M. and Le, Q. V. EfficientNet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning (ICML)*, 2019.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Wager, S., Wang, S., and Liang, P. S. Dropout training as adaptive regularization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

Zagoruyko, S. and Komodakis, N. Wide residual networks. In *British Machine Vision Conference (BMVC)*, 2016.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR), arXiv:1611.03530*, 2017a.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017b.

Chapter 5

Data augmentation and object representation in the brain

Contributors

Johannes Mehrer performed the representational similarity analysis. Nikolaus Kriegeskorte, Peter König and Tim C. Kietzmann reviewed and edited the manuscript submitted to CCN.

Outreach

This chapter extends the following publications:

- *Deep neural networks trained with heavier data augmentation learn features closer to representations in hIT.* **Alex Hernández-García**, Johannes Mehrer, Nikolaus Kriegeskorte, Peter König, Tim C. Kietzmann. Cognitive Computational Neuroscience (CCN), 2018.

One of the central goals of computational neuroscience is to develop better models of the human brain. The re-emergence of deep artificial neural networks, which now excel at many artificial intelligence tasks by automatically learning hierarchical representations (Girshick et al., 2014), has also had a positive impact on computational neuroscience. For instance, the features learnt by models trained for image object classification have been found to correlate better with the representations in the human inferior temporal cortex (hIT) than traditional hand-crafted features or shallow models (Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014; Güçlü & van Gerven, 2015). Further, convolutional neural networks are currently the most accurate models for multiple regions across the primate visual cortex (Kietzmann et al., 2019; Yamins & DiCarlo, 2016). However, while the similarity between artificial and biological neural networks is promising, a crucial question remains: what makes neural networks learn representations that more closely mirror activations in the brain?

Delving into this question is one of the goals of this thesis because of its potential implications on our understanding of the brain and learning systems in general. Previous work has revealed that networks performing better in classification tasks correlate more strongly with neural representations in high level areas (Yamins et al., 2014). However, the network architecture seems to play a crucial role (Storrs et al., 2017) and Mehrer et al. (2017) showed that training with more ecologically relevant image categories yields more similar representations. Inspired by the apparent importance of the training data, and the properties of data augmentation discussed in Chapter 4, we here explore the influence of data augmentation on the representational similarity between artificial neural networks and the human inferior temporal cortex.

As we have discussed in the Introduction (Chapter 1), the transformations included in (perceptually plausible) data augmentation schemes are inspired by the properties of visual percep-

tion. We perform translations, rotations, scaling and changes in the illumination of images (see Section 4.2.1) because these transformations are part of the variance we observe in the visual real-world. Transformations of this kind within certain ranges do not change the perceived object class and even identity. In Chapter 4 we have seen that applying these transformations to the training images of a neural network model is highly beneficial for generalisation. In this chapter we test the hypothesis that training with heavier data augmentation may encourage learning representations more aligned representations with the inferior temporal cortex.

5.1 Methods

This section presents the experimental setup to analyse the role of data augmentation on the similarity between artificial neural networks and neural representations in hIT. We describe the network architectures, the augmentation schemes and the methodology employed to compare both systems.

5.1.1 Network architectures

To increase the generality of our results, we analysed two distinct, well-known convolutional neural networks, which reach high-performance on image object-classification: the all convolutional network, All-CNN (Springenberg et al., 2014) and the wide residual network, WRN (Zagoruyko & Komodakis, 2016). We used the same architectures for the experiments in Chapter 4 and they are described in detail in Section 4.2.2, so we here only provide a brief overview of the most important properties:

- **All-CNN** consists only of 12 convolutional layers, each followed by batch normalisation and a ReLU activation. It has a total of 9.4 million parameters.
- **WRN** is a modification of ResNet (He et al., 2016) that achieves better performance with fewer layers, but more units per layer. We chose the WRN-28-10 version of the original paper, which has 28 layers and about 36.5 million parameters.

Following the conclusions from Chapter 4, we did not train the models with either weight decay or dropout, but we kept the rest of the hyperparameters as in the original papers.

5.1.2 Data augmentation

The goal of this work was to study the impact of data augmentation on the similarity of the representations with the activations in the inferior temporal cortex. For that purpose, we considered two data augmentation schemes: *light* and *heavier* augmentations, as described in Section 4.2.1. Below we summarise the transformations included in each scheme:

- The **light** augmentation scheme has been widely used in the literature, for instance (Springenberg et al., 2014). It performs only random horizontal flips and horizontal and vertical translations of maximum 10% of the image size. Additionally, we performed random crops of 128×128 pixels.
- The **heavier** scheme performs a larger range of random affine transformations such as scaling, rotations and shear mapping, as well as contrast and brightness adjustment and random crops.

We used these schemes to augment the highly benchmarked ImageNet ILSVRC 2012 data set (Russakovsky et al., 2015). We used ImageNet instead of CIFAR-10—for instance—because its higher resolution images more closely match the stimulus statistics of the human visual system. We resized the images into 150×200 pixels. Examples of the light and heavier augmentations on ImageNet photos are shown in Figure 5.1

Figure 5.1: Illustration of the transformations performed by the light and heavier augmentation schemes on two example images. Note that the five transformations of the images in this figure have been produced by setting extreme values of the parameters, so as to highlight the characteristics of the schemes and the differences between them.

The performance of All-CNN and WRN trained with light and heavier augmentation is shown in Figure 5.2. Note that training with light augmentation provides better results, specially on All-CNN. As pointed out in Chapter 4, this is likely explained by first, the fact that the heavier augmentation scheme was not designed to optimise classification, but rather as an arbitrary larger set of plausible transformations; and second, because the limited capacity of the models—especially All-CNN—may prevent them from exploiting the aggressive transformations of the already large ImageNet data set. Nonetheless, the objective of this study was to analyse the learnt representations given a reasonably accurate performance. Ideally, we would also analyse the representations of a model trained with no augmentation. However, the performance without data augmentation is significantly worse and this would likely impact the representations (Yamins et al., 2014).

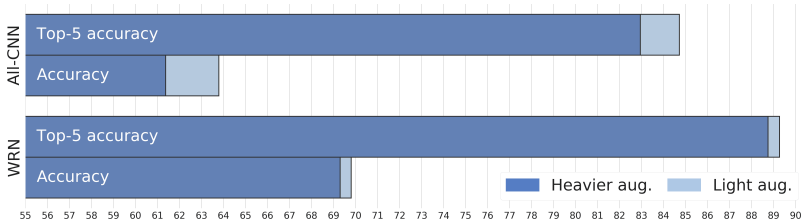


Figure 5.2: Test performance of All-CNN and WRN trained with light and heavier data augmentation.

5.1.3 Representational similarity analysis

In order to compare the representations learnt by the neural networks and the activations measured in the inferior temporal cortex, we made use of representational similarity analysis (RSA)

(Kriegeskorte et al., 2008a; Nili et al., 2014). The main advantage of RSA is that it allows direct comparisons across different model systems without having to explicitly align the different measurement types. This is accomplished by constructing representational dissimilarity matrices (RDMs) to express the pairwise similarity between stimuli, instead of directly comparing the representations of single stimuli. Across a set of input images, RDMs characterise the internal representations of a given system by storing all pairwise distances. The resulting matrix therefore expresses the representational geometry in the learnt activation space. By relying on distances, RDMs remain unchanged, if the space over which they are computed is rotated.

To characterise the representations in hIT, functional magnetic resonance imaging (fMRI) was used to measure BOLD responses while 15 participants were presented with 92 images of isolated objects. The images originate from a wide variety of categories and levels of abstraction. On the broadest level, they can be separated into animate and inanimate. Inanimate objects can either be natural or artificial, whereas animate objects are divided into human stimuli—heads and body parts—and animals—full body and heads only. This fMRI data set has been used in multiple studies and the details of the data acquisition can be found in (Kriegeskorte et al., 2008b). In Figure 5.3 we show the RDM of the brain data, and the RDMs of the WRN model for illustration. As in (Kriegeskorte et al., 2008b), in order to better visualise the differences across the RDM, the colour code represents the percentiles of the actual RDMs.

To compare artificial neural networks and hIT representations, the network activation profiles for the 92 images were extracted. In particular, we computed the activations at the outputs of the 12 ReLU layers of All-CNN and at the outputs of the residual blocks of WRN. We then computed the RDM of these activations using the Pearson correlation, as well as the RDM of the fMRI responses in hIT. To obtain a more compact representation of the CNN models, we combined the RDMs of all layers into a single RDM as a linear combination of the individual layer RDMs with respect to the hIT RDM using non-negative least squares and a cross-validation procedure, which avoids overfitting the image set.

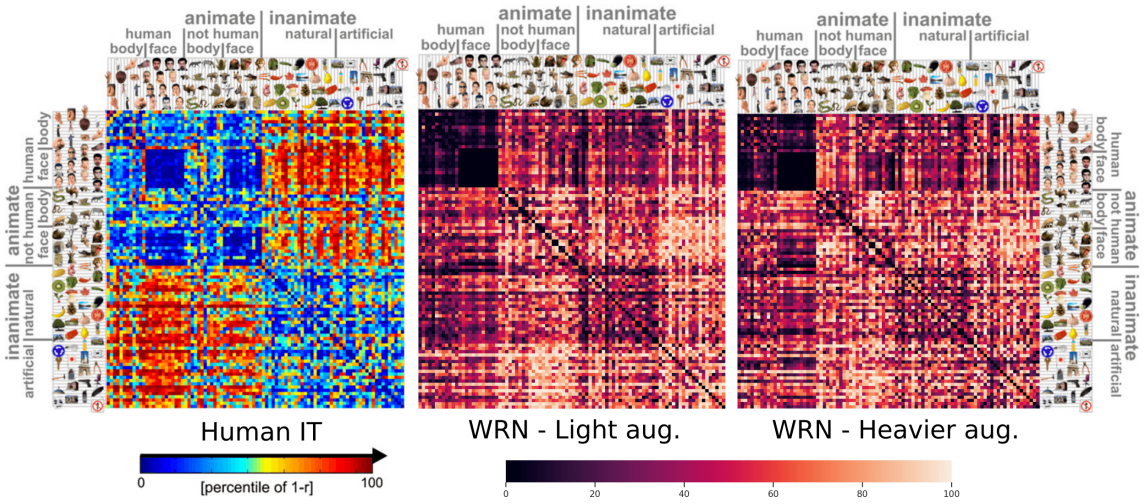


Figure 5.3: RDMs of the systems compared. Left: adapted from (Kriegeskorte et al., 2008b), the RDM of the human inferior temporal cortex. Centre and right, the RDMs of the WRN model, trained with light and heavier augmentation, respectively. As in (Kriegeskorte et al., 2008b), the colour code in the matrices shown in the figure represents the percentile of the dissimilarity.

Finally, we characterise the similarity between the artificial neural networks and hIT by computing the Kendall's rank correlation coefficient τ_A between the RDM of the hIT representations and the RDM of the convolutional models. Standard errors were obtained from the similarity estimates of the 15 human subjects.

5.2 Results and discussion

We show the results of the representational similarity analysis to compare the representations learnt by the neural networks and the fMRI data in Figure 5.4. As a main conclusion, we found that the correlation with the hIT representations is significantly higher for the models trained with heavier data augmentation. Not only is this indicated by the Kendall correlation, but also by visual inspection, the RDM of the model trained with heavier augmentations seems more similar to the RDM of the human IT. For example, face images—both human and non-human—form clearer similarity clusters in the model trained with heavier data augmentation, which is a well-studied property of the primate visual cortex.

In the case of the wide residual network (WRN) the difference between the two levels of augmentation is considerably larger, while in the All-CNN models, although statistically significant ($p < 0.05$), the difference is smaller. However, recall that the classification performance of the models trained with heavier augmentations is worse, especially in the case of All-CNN (Figure-5.2. Therefore, it seems that even though the more aggressive transformations do not improve the classification performance, they do increase the similarity with the inferior temporal cortex.

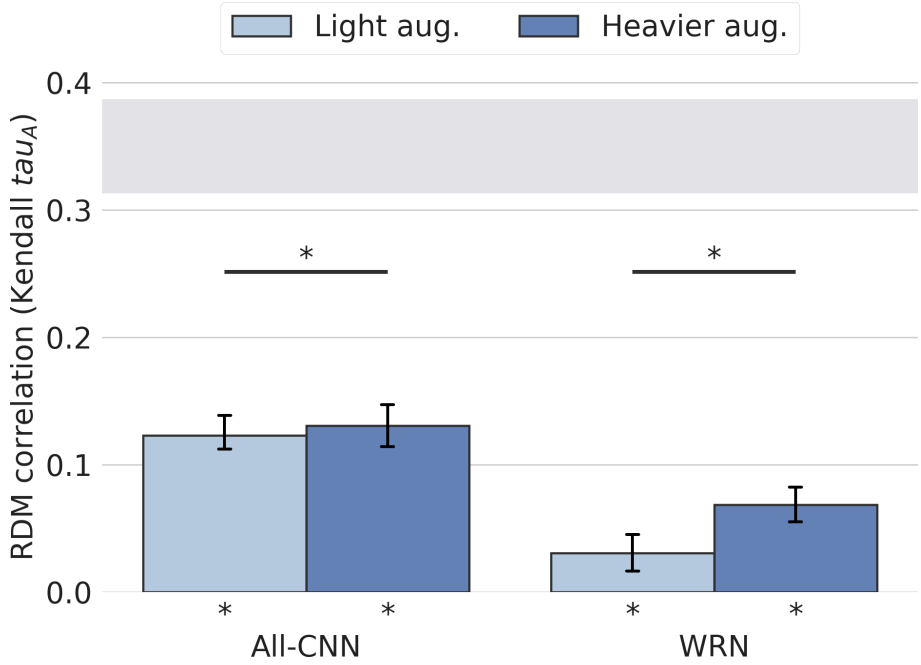


Figure 5.4: Comparison of the Kendall's τ_A coefficient of the hIT RDM and the RDM of the networks trained with light and heavier data augmentation. Both on All-CNN and WRN, the correlation of the model trained with heavier transformations is significantly higher than the light counterpart. The grey shaded area indicates the maximum possible correlation of a model given the noise in the measured data.

It is also interesting that the similarity with hIT is higher for All-CNN, again despite the lower performance, as opposed to previous work that indicated a correlation between performance and similarity with the visual cortex (Yamins et al., 2014). The overall lower correlation of WRN adds more evidence to the conclusion of Storrs et al. (2017), who showed that residual networks exhibit a particularly low correlation with hIT compared to other architectures.

Given the exploratory nature of this study, it is not yet clear what exact mechanisms lead to the better match between representational geometries in higher level visual cortex and networks trained with heavier data augmentation. One hypothesis is that the larger variety during training may be more biologically plausible than training with constant images or very light transformations. Humans develop robust object representations based on highly variable input, while freely

exploring the world. Sources of variation include different orientations, lighting conditions, backgrounds and occlusion. Eye-movements, including drifts and *microsaccades*, may further contribute to the variability in the sensory input to which the recognition has to be robust. As we will further discuss in Chapter 6, this robustness is reflected in invariant activations towards identity-preserving transformations in the higher visual cortex.

Our experiments addressed the question as to which factors drive computational models to learn features closer to the brain representations. Given the superiority in visual robustness of the human brain, these insights may have implications for artificial vision systems based on deep neural networks, and for ANNs as a model system for visual processing in the brain. Finding that heavier data transformations leads to more IT-like representations further supports the notion that the input distribution plays a crucial role during the learning of representations in both the brain artificial networks.

5.3 Conclusions

In this chapter, we have explored how far light and heavier augmentation of the training set can affect the internal representations of deep neural networks and their alignment with human IT. To compare the neural and model system, we used representational similarity analysis, which allows for straightforward comparisons across different modalities—in this case, fMRI BOLD signal and neural network activations. RSA revealed that the neural networks trained with heavier transformations learn representations more similar to those observed in higher visual cortex.

Future work should analyse a larger range of network architectures and data sets to gain better insights into the mechanisms driving the internal representations. It will be also interesting to study the different components of data augmentation in order to understand which particular transformations play a bigger role in better explaining hIT.

Bibliography

Girshick, R., Donahue, J., Darrell, T., and Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Güçlü, U. and van Gerven, M. A. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 2015.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Khaligh-Razavi, S.-M. and Kriegeskorte, N. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLOS Computational Biology*, 2014.

Kietzmann, T. C., McClure, P., and Kriegeskorte, N. Deep neural networks in computational neuroscience. *Oxford Research Encyclopedia of Neuroscience*, 2019.

Kriegeskorte, N., Mur, M., and Bandettini, P. A. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2008a.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., and Bandettini, P. A. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 2008b.

Mehrer, J., Kietzmann, T. C., and Kriegeskorte, N. Deep neural networks trained on ecologically relevant categories better explain human IT. In *Conference on Cognitive Computational Neuroscience (CCN)*, 2017.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., and Kriegeskorte, N. A toolbox for representational similarity analysis. *PLOS Computational Biology*, 2014.

Russakovsky, O. et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 2015.

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. Striving for simplicity: The all convolutional net. In *International Conference on Learning Representations (ICLR), arXiv:1412.6806*, 2014.

Storrs, K., Mehrer, J., Walther, A., and Kriegeskorte, N. Architecture matters: How well neural networks explain IT representation does not depend on depth and performance alone. In *Conference on Cognitive Computational Neuroscience (CCN)*, 2017.

Yamins, D. L. and DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 2016.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences (PNAS)*, 2014.

Zagoruyko, S. and Komodakis, N. Wide residual networks. In *British Machine Vision Conference (BMVC)*, 2016.

Chapter 6

Data augmentation invariance

Contributors

Tim C. Kietzmann supervised the work during my internship at his lab. Tim and Peter König reviewed and edited the manuscripts submitted to conferences.

Outreach

This chapter extends the following publications:

- *Learning robust visual representations using data augmentation invariance.* **Alex Hernández-García**, Peter König, Tim C. Kietzmann. arXiv preprint arXiv:1906.04547 & Cognitive Computational Neuroscience (CCN), 2019 & Workshop on Bridging AI and Cognitive Science, International Conference on Learning Representations (ICLR), 2019.
- *Learning representational invariance instead of categorization.* **Alex Hernández-García**, Peter König. Workshop on pre-registration in computer vision, International Conference on Computer Vision (ICCV), 2019.

Deep artificial neural networks (ANNs) have borrowed much inspiration from neuroscience and are, at the same time, the current best model class for predicting neural responses across the visual system in the brain (Kietzmann et al., 2019; Kubilius et al., 2018). Yet, despite consensus about the benefits of a closer integration of deep learning and neuroscience (Bengio et al., 2015; Marblestone et al., 2016; Richards et al., 2019), important differences remain (Sinz et al., 2019; Geirhos et al., 2020; Dujmović et al., 2020).

Here, we investigate a representational property that is well established in the neuroscience literature on the primate visual system: the increasing robustness of neural responses to identity-preserving image transformations. While early areas of the ventral stream (V1-V2) are strongly affected by variation in object size, position, viewpoint and other factors, later levels of processing are increasingly robust to such changes, as measured first in single neurons of the inferior temporal (IT) cortex of macaques Booth & Rolls (1998) and then in humans' (Quiroga et al., 2005; Isik et al., 2013). The cascaded achievement of invariance to such identity-preserving transformations has been proposed as a key mechanism for robust object recognition (DiCarlo & Cox, 2007; Tacchetti et al., 2018).

Learning such invariant representations has been a desired objective since the early days of artificial neural networks (Simard et al., 1992). Accordingly, a myriad of techniques have been proposed to attempt to achieve tolerance to different types of transformations (Cohen & Welling (2016) briefly reviewed some of these efforts). Interestingly, recent theoretical work (Achille & Soatto, 2018) has shown that invariance to "nuisance factors" should naturally emerge from the optimisation process of deep models that minimise the information of the representations about

the inputs, while retaining the minimum information about the target, as proposed by Tishby & Zaslavsky (2015) in the information bottleneck principle.

Nevertheless, artificial neural networks are still not robust to identity-preserving transformations, including simple image translations (Zhang, 2019). A remarkable extreme example are adversarial attacks (Szegedy et al., 2013), in which small changes, imperceptible to the human brain, can alter the classification output of the network. Extending this line of research, we used data augmentation as a framework to generate the transformations to which vision models should be invariant to, according to visual perception and biological vision. As a first contribution, we propose a simple metric, *data augmentation invariance score*, to measure the invariance of neural networks to identity-preserving transformations.

Second, inspired by the increasing invariance observed along the primate ventral stream of the visual cortex, we here propose *data augmentation invariance*, a simple, yet effective and efficient mechanism to improve the robustness of the representations: we include an additional contrastive term in the objective function that encourages the similarity between augmented examples within each batch. We will argue that this objective encodes a useful inductive bias that exploits prior knowledge from visual perception and biological vision.

Finally, we explore the possibility of fully replacing the categorisation objective that is commonly used to train neural networks for classification, the categorical cross-entropy, by objective functions purely based on invariance learning.

6.1 Invariance score

To measure the invariance of the learnt features under the influence of identity-preserving image transformations we compare the activations of a given image with the activations of an augmented version of the same image.

Consider the activations of an input image x at layer l of a neural network, which can be described by a function $f^{(l)}(x) \in \mathbb{R}^{D^{(l)}}$. We can define the distance between the activations of two input images x_i and x_j by their mean squared difference:

$$d^{(l)}(x_i, x_j) = \frac{1}{D^{(l)}} \sum_{k=1}^{D^{(l)}} (f_k^{(l)}(x_i) - f_k^{(l)}(x_j))^2 \tag{6.1}$$

Following this, we are interested in the mean squared distance between $f^{(l)}(x_i)$ and a randomly sampled transformation of x_i, that is $d^{(l)}(x_i, G(x_i))$. $G(x)$ refers to the stochastic function that transforms the input images according to a pre-defined, parameterised data augmentation scheme.

In order to assess the similarity between the activations of an image x_i and of its augmented versions $G(x_i)$ we normalise it by the average similarity in the (test) set. We define the *data augmentation invariance score* $S_i^{(l)}$ of image x_i towards the transformation $G(x)$ at layer l of a model, with respect to a data set of size N, as follows:

$$S_i^{(l)} = 1 - \frac{d^{(l)}(x_i, G(x_i))}{\frac{1}{N} \sum_{j=1}^{N} d^{(l)}(x_i, x_j)} \tag{6.2}$$

Note that the invariance $S_i^{(l)}$ takes the maximum value of 1 if the activations of x_i and its transformed version $G(x_i)$ are identical, and the value of 0 if the distance between transformed examples, $d^{(l)}(x_i, G(x_i))$ (numerator), is equal to the average distance in the set (denominator).

6.1.1 Learning objective

Most ANNs trained for object categorisation are optimised through mini-batch stochastic gradient descent (SGD), that is the weights are updated iteratively by computing the loss of a batch $\mathcal{B}$ of examples, instead of the whole data set at once. The models are typically trained for a number of *epochs* E which is a whole pass through the entire training data set of size N. That is, the weights are updated $K = \frac{N}{|\mathcal{B}|}$ times each epoch.

Data augmentation introduces variability into the process by performing a different, stochastic transformation of the data every time an example is fed into the network. However, with standard data augmentation, the model has no information about the *identity* of the images. In other words, different augmented examples, seen at different epochs, separated by $\frac{N}{|\mathcal{B}|}$ iterations on average, correspond to the same seed data point. This information is potentially valuable and useful to learn better representations. For example, in biological vision, the high temporal correlation of the stimuli that reach the visual cortex may play a crucial role in the creation of robust connections (Becker, 1999; Kording et al., 2004; Wyss et al., 2006). However, this is generally not exploited in supervised settings. In semi-supervised learning, where the focus is on learning from fewer labelled examples, data augmentation has been used as a source of variability together with dropout or random pooling, among others (Laine & Aila, 2016).

In order to make use of this information and improve the robustness, we first propose *in-batch* data augmentation by constructing the batches with M transformations of each example—Hoffer et al. (2019) recently discussed a similar idea. Additionally, we propose a new objective function that accounts for the invariance of the feature maps across multiple image samples. Considering the difference between the activations at layer l of two images, $d^{(l)}(x_i, x_j)$, defined in Equation 6.1, we define the *data augmentation invariance* loss at layer l for a given batch $\mathcal{B}$ as follows:

$$\mathcal{L}_{inv}^{(l)} = \frac{\sum_k \frac{1}{|\mathcal{S}_k|^2} \sum_{x_i, x_j \in \mathcal{S}_k} d^{(l)}(x_i, x_j)}{\frac{1}{|\mathcal{B}|^2} \sum_{x_i, x_j \in \mathcal{B}} d^{(l)}(x_i, x_j)} \tag{6.3}$$

where $\mathcal{S}_k$ is the set of samples in the batch $\mathcal{B}$ that are augmented versions of the same seed sample x_k. This loss term intuitively represents the average difference of the activations between the sample pairs that correspond to the same source image, relative to the average difference of all pairs. A convenient property of this definition is that $\mathcal{L}_{inv}$ does not depend on either the batch size or the number of in-batch augmentations $M = |\mathcal{S}_k|$. Furthermore, it can be efficiently implemented using matrix operations. Our data augmentation invariance can be seen as a contrastive loss (Hadsell et al., 2006), since the aim is to bring closer the representations of related examples—transformations of the same source image—and push apart the representations from other examples.

In order to simultaneously achieve certain representational invariance at L layers of the network and high object recognition performance at the network output, we define the total loss as follows:

$$\mathcal{L} = (1 - \alpha)\mathcal{L}_{obj} + \sum_{l=1}^{L} \alpha^{(l)} \mathcal{L}_{inv}^{(l)} \tag{6.4}$$

where $\alpha = \sum_{l=1}^{L} \alpha^{(l)}$ and $\mathcal{L}_{obj}$ is the loss associated with the object recognition objective, typically the cross-entropy between the object labels and the output of a softmax layer. For all the experiments presented in this chapter, we set $\alpha = 0.1$ and distributed the coefficients across the layers according to an exponential law, such that $\alpha^{(l=L)} = 10\alpha^{(l=1)}$. This aims at simulating a probable response along the ventral visual stream, where higher regions are more invariant than the early visual cortex[1].

[1]It is beyond the scope of this study to analyse the sensitivity of the hyperparameters $\alpha^{(l)}$, but we did not observe a significant impact in the classification performance by using other distributions.

6.1.2 Architectures and data sets

As test beds for our hypotheses and proposal we trained three neural network architectures: all convolutional network, All-CNN-C (Springenberg et al., 2014); wide residual network, WRN-28-10 (Zagoruyko & Komodakis, 2016); and DenseNet-BC (Huang et al., 2017). All three architectures have been widely used in the deep learning literature, including our experiments in Chapter 4, where the details can be consulted. They provide generality of the results, as they have distinctive architectural characteristics: only convolutional layers, residual blocks and dense connectivity, respectively.

We trained the three architectures on the highly benchmarked data set for object recognition CIFAR-10 (Krizhevsky & Hinton, 2009). Additionally, in order to test our proposal on higher resolution images and a larger number of classes, we also trained All-CNN and WRN on the *tiny* ImageNet data set, a subset of ImageNet (Russakovsky et al., 2015) with 100,000 64x64 training images that belong to 200 classes. All models were trained using the *heavier* data augmentation scheme as defined in Section 4.2.1, which consists of affine transformations, contrast adjustment and brightness adjustment.

For the models trained on CIFAR-10, the training hyperparameters—learning rate, decay, number of epochs, etc.—were set as in the original papers, except that, following the conclusions from Chapter 4, we did not use explicit regularisation—weight decay and dropout—since comparable performance is obtained without them if data augmentation is used. For all three architectures, we performed $M = |\mathcal{S}_k| = 8$ augmentations per image in the batch, while keeping the real batch size as in the original papers.

On tiny ImageNet, All-CNN included three additional layers and was trained for 150 epochs, with batch size 128 and initial learning rate 0.01 decayed by 0.1 at epochs 100 and 125. We included $M = 8$ augmentations per image in the batches. WRN was trained for 50 epochs, with batch size 32 and initial learning rate 0.01 decayed by 0.2 at epochs 30 and 40. To train WRN with data augmentation invariance, we performed $M = 4$ augmentations per image. In all cases, the models were trained with stochastic gradient descent with Nesterov momentum 0.9.

Note that the hyperparameters were fine-tuned for the original papers by training only with the standard categorical cross-entropy and with standard epoch-wise data augmentation. Therefore, they were likely suboptimal for our proposed data augmentation invariance. However, our aim was not achieving the best possible classification performance, but rather demonstrate the suitability of data augmentation invariance and analyse the learnt representations.

The invariance loss defined in Equation 6.3 was computed after the ReLU activation of each convolutional layer for All-CNN, at the output of each residual block for WRN, and after the first convolution and the output of each dense block for DenseNet.

6.2 Results

The first contribution of this chapter is to empirically test in how far convolutional neural networks trained for object categorisation produce invariant representations under the influence of identity-preserving transformations of the input images. Figures 6.1–6.3 show the invariance scores, as defined in Equation 6.2, across the network layers. Since we do not compute the invariance score at every single layer of the architectures, the numbering of the layers simply indicate increasing depth in the hierarchy (see Section 6.1.2 for the details). The red boxes correspond to the baseline models and the blue boxes to the models trained with our data augmentation invariance objective.

The distributions of the invariance score shown in the figures were computed using the test partitions of the data. For each image, we performed five random transformations using as parameters the values at exactly half of the range used for training (see Section 4.2.1). Despite the presence of data augmentation during training, which implies that the networks *sees* and may learn augmentation-invariant transformations, the representations of the baseline models (red

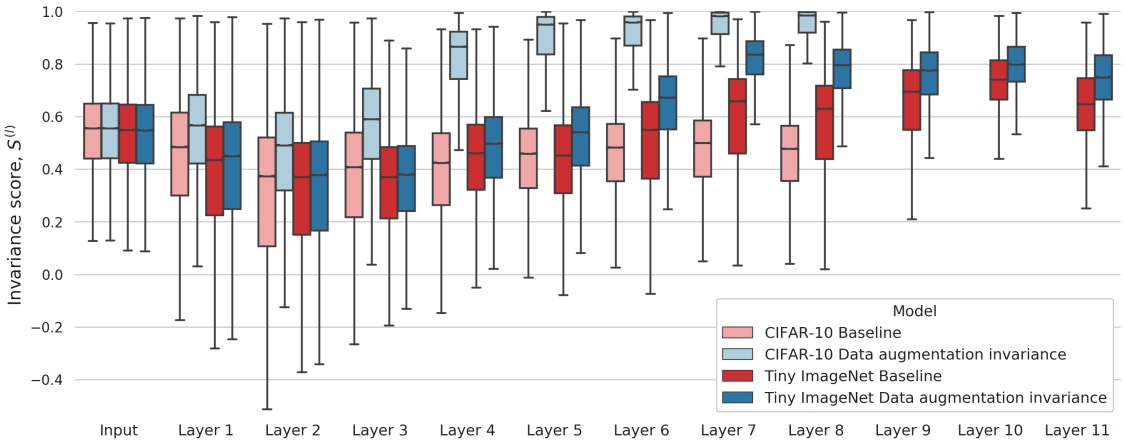


Figure 6.1: All-CNN: distribution of the invariance score at each layer of the baseline model and the model trained data augmentation invariance (higher is better).

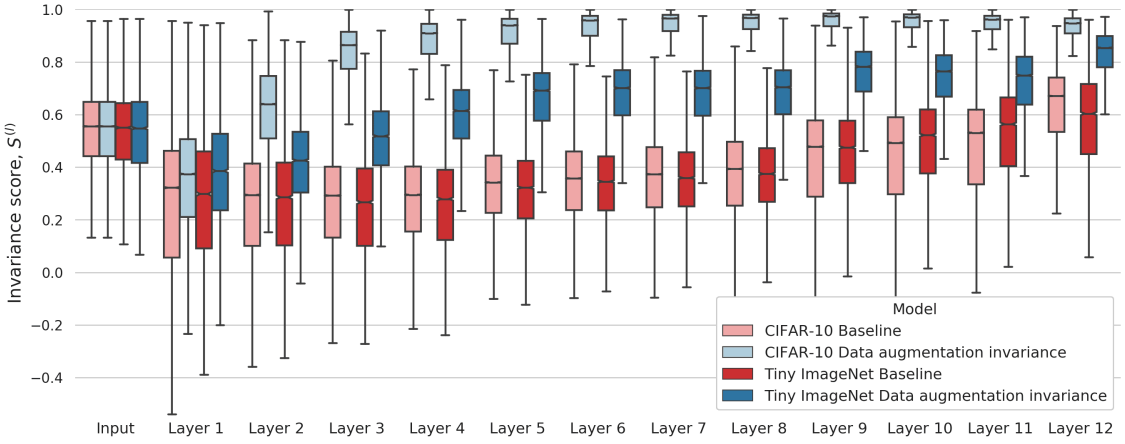


Figure 6.2: WRN: distribution of invariance score at each layer of the baseline model and the model trained data augmentation invariance (higher is better).

boxes) do not increase substantially beyond the invariance observed in the pixel space (left-most boxes). To illustrate this, see the images in Figure 6.4, whose representations are all equally distant to the reference image, despite the perceptual similarity of the transformed images.

As a solution, we have proposed a simple, label-free modification of the loss function to encourage the learning of data augmentation-invariant features. The blue boxes in Figures 6.1–6.3 show that our invariance mechanism pushed network representations to become increasingly more robust with network depth[2]. As discussed above, this is a well-studied property of the visual ventral stream in the primate brain.

In order to better understand the effect of the data augmentation invariance objective, we analysed the learning dynamics of the invariance loss at each layer of All-CNN trained on CIFAR-10. In Figure 6.5, we see that in the baseline model, the invariance loss keeps increasing over the course of training. In contrast, in the models trained with data augmentation invariance, the loss drops for all but the last layer. Perhaps unexpectedly, the invariance loss increases during the first epochs and only then starts to decrease. While further investigation is required, these two phases may correspond to the fitting and compression-diffusion phases proposed in the framework of

[2]Both All-CNN and WRN seem to more easily achieve the representational invariance on CIFAR-10 than on Tiny ImageNet. This may indicate that the complexity of the data set not only makes the object categorisation more challenging, but also the learning of invariant features.

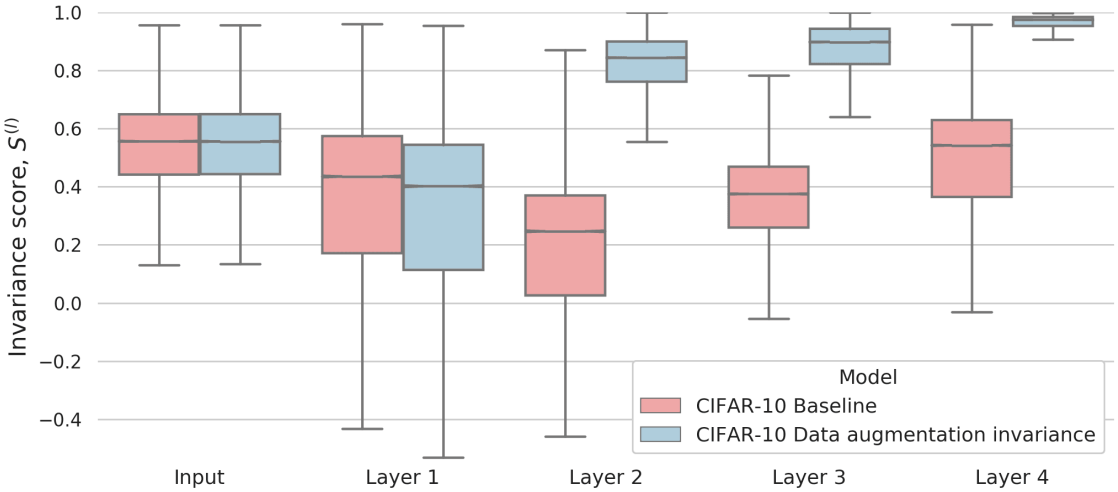


Figure 6.3: DenseNet: distribution of invariance score at each layer of the baseline model and the model trained data augmentation invariance (higher is better).

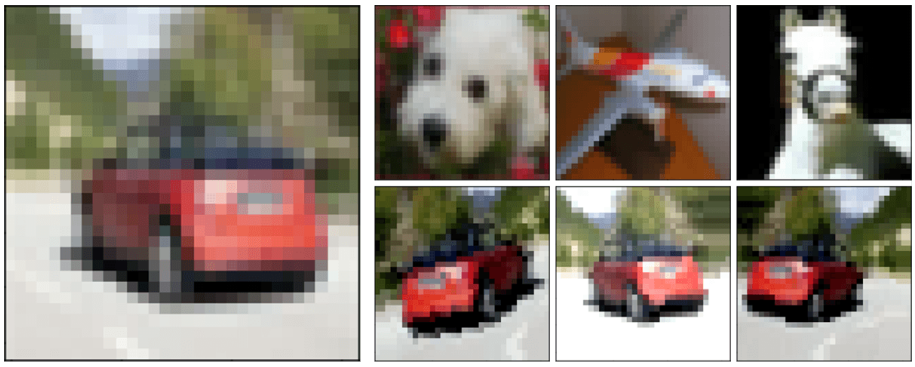


Figure 6.4: The top layer representations of the six images on the right learnt by All-CNN are equally (dis)similar to the reference image (left), even though the images at the bottom row are transformations of it.

the information bottleneck principle by Shwartz-Ziv & Tishby (2017). According to the authors, during the first epochs of optimisation with SGD, the model increases the information about the labels (fitting) and during the rest of training, the model reduces the information on the input (compression). However, note that Saxe et al. (2019) have argued that this occurs only in some cases that depend on the non-linearities.

In terms of efficiency, adding terms to the objective function implies an overhead of the computations. However, since the pairwise distances can be efficiently computed at each batch through matrix operations, the training time was only increased by about 10 % on average.

Finally, as reported in Table 6.1, the improved invariance comes at little or no cost in the categorisation performance, as the networks trained with data augmentation invariance achieved similar classification performance to the baseline model and in some cases it clearly improved it. This is remarkable as the hyperparameters used in all cases were optimised to maximise the performance in the original models, trained without data augmentation invariance. Therefore, it is reasonable to expect an improvement in the classification performance if, for instance, the batch size or the learning rate schedule are better tuned for this new learning objective. Learning increasingly invariant features could lead to more robust categorisation, as exemplified by the increased test performance observed for the All-CNN models—despite no hyperparameter tuning.

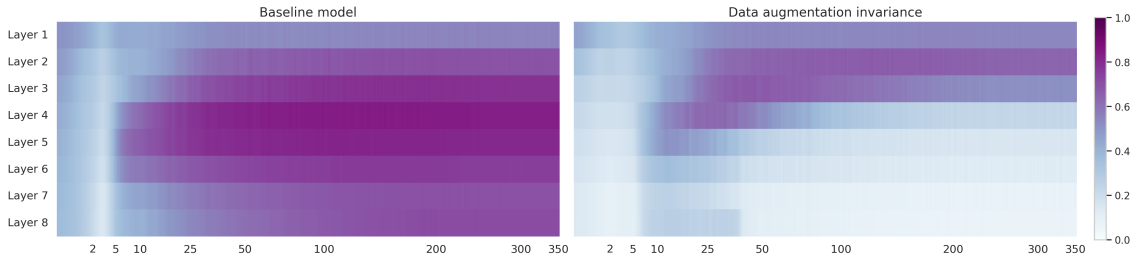


Figure 6.5: Dynamics of the data augmentation invariance loss $\mathcal{L}_{inv}^{(l)}$ during training (lower is better). The axis of abscissas (epochs) is scaled quadratically to better appreciate the dynamics at the first epochs. The same random initialisation was used for both models.

	CIFAR-10			Tiny ImageNet (acc. — top5)	
	All-CNN	WRN	DenseNet	All-CNN	WRN
Baseline	91.48	94.58	94.88	51.09 — 73.48	61.49 — 82.99
Data aug. invariance	92.47	94.86	93.98	52.57 — 76.53	61.23 — 83.23

Table 6.1: Classification accuracy on the test set of the baseline models and the models trained with data augmentation invariance.

6.3 Learning representational invariance *instead* of categorisation

In view of the effectivity of the data augmentation invariance learning objective, it is reasonable to wonder whether such an objective could fully replace the standard categorisation objective commonly used in so-called[3] supervised learning. There are multiple reasons why exploring alternatives to classification objectives is attractive. In the Introduction of this thesis we have discussed the benefits of incorporating inductive biases from visual perception and biological vision in the form of objective functions and the results presented above in this chapter have demonstrated the usefulness of data augmentation to promote invariant representations.

A remarkable example of the mismatch between ANNs and primate visual perception is the well-known vulnerability of the former to adversarial perturbations (Szegedy et al., 2013; Dujmović et al., 2020). Recent work by Ilyas et al. (2019) has suggested that adversarial vulnerability might be caused by highly discriminative features present in the data yet incomprehensible to humans. In the same line, it has been recently shown Wang et al. (2019) that ANNs learn high-frequency components of images, imperceptible to humans, but useful for categorisation. A related idea was suggested earlier by Jo & Bengio (2017). Notably, this is only one example of the differences between current artificial and biological visual object perception (Sinz et al., 2019; Geirhos et al., 2020; Malhotra et al., 2020). We hypothesise that some of these undesired properties might be caused by the optimisation of the specific task of classification.

In this section we present the results of a exploratory, preliminary study where we aim to replace the standard categorical-cross entropy objective by a combination of data augmentation invariance and a similarly defined *class-wise invariance*, which uses the labels of the images.

6.3.1 Class-wise invariance

Class-wise invariant representation learning Belharbi et al. (2017) was introduced as a regularisation term that encourages similarity in the representations of objects from the same class. The authors showed that class-wise invariance helps improve generalisation, especially when few ex-

[3]See the discussion about supervised learning in the Introduction (Section 1.3)

amples are available. Related ideas have been proposed in the metric learning and clustering literature.

Class-wise invariance is interesting because, in spite of simply optimising the prediction of the object labels, it sets the learning objective on how the intermediate features should be like. However, used on its own it would possibly be subject to some of the same undesirable properties of purely supervised methods. We hypothesise that combined, data augmentation and class-wise invariance alone—without a categorisation objective—may learn robust, discriminative features.

We define the class-wise invariance loss at layer l of a neural network $\mathcal{L}_C^{(l)}$ as a parallel to the data augmentation invariance loss (Equation 6.3):

$$\mathcal{L}_C^{(l)} = \frac{\sum_r \frac{1}{|\mathcal{T}_r|^2} \sum_{x_i, x_j \in \mathcal{T}_r} d^{(l)}(x_i, x_j)}{\frac{1}{|\mathcal{B}|^2} \sum_{x_i, x_j \in \mathcal{B}} d^{(l)}(x_i, x_j)} \tag{6.5}$$

where $\mathcal{T}_r$ are the subsets from $\mathcal{B}$ formed by images of the same object class r. Let us denote by $\mathcal{L}_{DA}^{(l)}$ the data augmentation invariance loss (Equation 6.3). We propose to optimise, through stochastic gradient descent, the following overall objective:

$$\mathcal{L} = \sum_{l=1}^{L} \alpha^{(l)} \mathcal{L}_{DA}^{(l)} + \sum_{l=1}^{L} \beta^{(l)} \mathcal{L}_C^{(l)} \tag{6.6}$$

where $\alpha^{(l)}$ and $\beta^{(l)}$ are scalars that control the degree of similarity between the features of augmented samples and of objects of the same category, respectively, at each layer l of the architecture. Summarised, by jointly optimising $\mathcal{L}_{DA}^{(l)}$ and $\mathcal{L}_C^{(l)}$, we expect the model to learn robust features—as encouraged by the data augmentation invariance—while still allowing for categorisation—driven by the class-wise invariance.

6.3.2 Results

In order to test this idea, we trained All-CNN on CIFAR-10 with the objective defined in Equation 6.6. We found that optimising this objective function as is, with the hyperparameters as in the models trained with standard categorical cross-entropy was *not* able to perform multi-class classification. One hypothesis is that the learnt representations collapse along one single dimension, since the explained variance by the first principal component gets close to 100 % and, therefore, the data points get separated in two clusters only. We have not found yet the exact cause of the undesired behaviour, but it may be due to the inability to escape local minima.

Despite the limitation of the fact that a 10-classes problem could not be optimised, in order to gain insights on the effect of the method, we analysed the representations learnt through the optimisation of invariance, instead of categorisation. In Figure 6.6, we plot the representational dissimilarity matrices of the invariance model, alongside the baseline model trained with categorical cross-entropy and an additional model that jointly optimised both objectives. The latter model did achieved test performance on CIFAR-10 comparable to the baseline model.

Interestingly, we found that the models optimised with invariance objectives learnt representations that naturally create meaningful clusters that separate animals and vehicle classes, that is animate and inanimate objects. This categorisation has been reported multiple times to be distinctively represented in the inferior temporal cortex of the primate brain (Kriegeskorte et al., 2008; Bao et al., 2020). This connection is still speculative and in the future we will further explore the representations learnt through this invariance objectives.

In an additional effort to further understand the learnt representations without the limitation of the low accuracy on 10 classes, we created a subset of CIFAR-10 with the images of cars and dogs only, that is a binary classification problem. This data set could be classified correctly with high accuracy. In Figure 6.7, we see that in the model trained with the invariance objective most

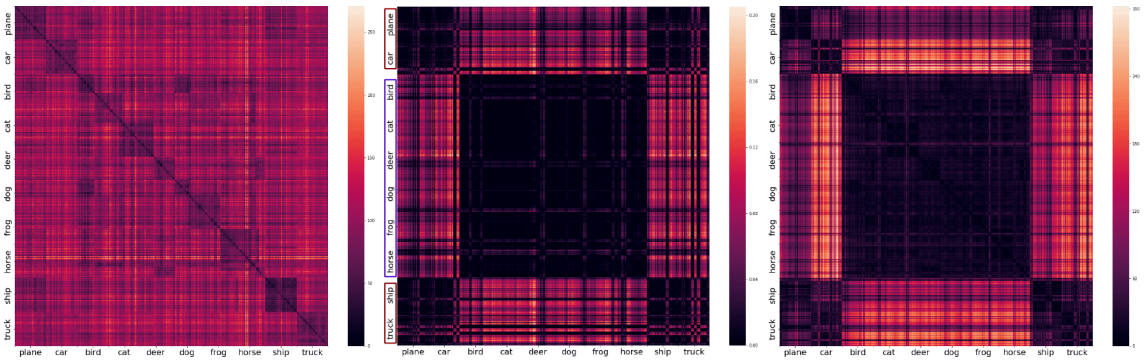
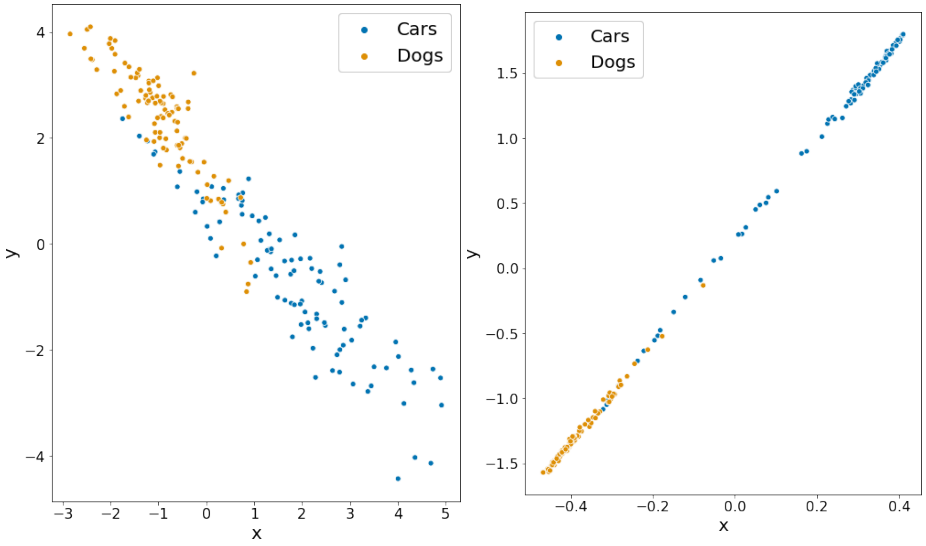


Figure 6.6: Representational similarity matrices of All-CNN trained on CIFAR-10 with (left) standard categorical cross-entropy, (middle) the invariance loss defined in Equation 6.6 and (right) the invariance loss plus a categorical-cross entropy term. In the models trained with invariance losses, meaningful hierarchical clusters (animals vs. vehicles) emerge.

examples of the two classes are separated by a larger margin than in the baseline model, where the clusters of the two classes are closer to each other.



(a) Baseline model

(b) Invariance model

Figure 6.7: Representations of the test *CIFAR-2* (cars and dogs) examples along the two principal components of the top layer representations of All-CNN.

Finally, we evaluated the adversarial robustness of the models. We had hypothesised that one of the reasons for adversarial vulnerability is that models optimised for categorisation only are highly unconstrained, with no specification of how the features should be. Therefore, the model focuses on learning any features that allows for linearly separable classes at the output of the network (Malhotra et al., 2020). This has been shown to be prone to adversarial vulnerability, that is small perturbations largely affect the classification. In contrast, our invariance objective can be seen as the opposite strategy. It specifies how the features should be—increasingly invariant to identity-preserving transformations and relatively invariant within object classes—and expects good classification as a by-product. This could improve the sensitivity to adversarial perturbations and is what our results in Table 6.2 suggest: the model trained with the invariance objectives is remarkably more robust that the baseline models, without sacrificing performance.

	Baseline (only cat.)	Only invariance
Clean examples	98.9 %	98.7 %
Attack: PGD $\varepsilon = 0.03$	7.6 %	62.0 %
Attack: FGSM $\varepsilon = 0.03$	31.5 %	88.5 %

Table 6.2: Adversarial robustness of the baseline (only categorisation) and invariance models trained on *CIFAR-2*. Without any adversarial training, the model trained with the invariance objective only is highly robust to adversarial perturbations, in contrast to the standard, categorisation model.

6.4 Discussion

In this chapter, we have first proposed an invariance score that assesses the robustness of the features learnt by a neural network towards the identity-preserving transformations typical of common data augmentation schemes (see Equation 6.2). Intuitively, the more similar the representations of transformations of the same image, with respect to other images in a data set, the higher the data augmentation invariance score. The data augmentation score is meaningful in that the transformations performed by perceptually plausible data augmentation schemes—which we consider here exclusively—are motivated by human visual perception and coincide largely with the transformations to which the higher visual cortex is invariant.

Using this score, we have analysed the representations learnt by three popular neural network architectures—All-CNN, WRN and DenseNet—trained on image object recognition tasks—CIFAR-10 and Tiny ImageNet. The analysis revealed that their features are less invariant than commonly assumed, despite sufficient exposure to matching image transformations during training. In most cases, the representational invariance did not even increase with respect to the original pixel space. This property is fundamentally different to the primate ventral visual stream, where neural populations have been found to be increasingly robust to changes in view or lighting conditions of the same object (DiCarlo & Cox, 2007).

Taking inspiration from this property of the visual cortex, we have proposed a label-free objective to encourage learning more robust features, using data augmentation as the framework to perform identity-preserving transformations on the input data. We constructed mini-batches with M augmented versions of each image and modified the loss function to maximise the similarity between the activations of the same seed images, as compared to other images in the batch. Aiming to approximate the observations in the biological visual system, higher layers were set to achieve exponentially more invariance than the early layers. An interesting avenue for future work will be to investigate whether this increased robustness also allows for better modelling of neural data.

Data augmentation invariance effectively produced more robust representations, unlike standard models optimised only for object categorisation, at little or no cost in classification performance and with an affordable, slight increase (10 %) in training time. Ideally, object recognition models should be reasonably invariant to all the transformations of the objects to which human perception is also invariant. Data augmentation is just an approximation to analyse and encourage invariance to a set of transformations that can be applied on still, 2D images. Future work should analyse the invariance of models trained with video (Taylor et al., 2010) and even 3D data (Achlioptas et al., 2018).

These results provide additional empirical evidence that deep supervised models optimised only according to the standard categorisation objective—the categorical cross-entropy between the true object labels and the model predictions—learn discriminative but non-robust features. This is likely due to their large capacity to learn discriminative features in too unconstrained a setting (Geirhos et al., 2020; Malhotra et al., 2020), which has been recently suggested to be at the source of adversarial vulnerability (Ilyas et al., 2019).

In order to explore whether getting rid of the categorisation objective can produce features

more aligned with our intuitions from visual perception, less vulnerable to adversarial perturbations and potentially more similar to the brain representations, we proposed to replace the categorical cross-entropy with purely invariance objectives. In particular, we combined layer-wise data augmentation invariance with class-wise invariance, which encourages similarity of the features from images of the same class. Although our preliminary experiments did not succeed at achieving competitive categorisation performance, we made interesting observations in our analysis.

First, training with invariance objectives yields representations that are clustered hierarchically, while the dissimilarity matrix of the standard model is fairly homogeneous. Remarkably, the main clusters formed through invariance learning correspond to the animate and inanimate classes, a separation consistently observed in the primate visual cortex Kriegeskorte et al. (2008); Bao et al. (2020). Furthermore, we trained a model on a binary classification task using invariance objectives only and found that the adversarial vulnerability is very low, in contrast to categorisation models, which exhibit very high sensitivity to adversarial perturbations. While these conclusions are still speculative, they set a promising path for future research on alternative objective functions that encode inductive biases from visual perception and biological vision.

Bibliography

Achille, A. and Soatto, S. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research (JMLR)*, 2018.

Achlioptas, P., Diamanti, O., Mitliagkas, I., and Guibas, L. Learning representations and generative models for 3D point clouds. In *International Conference on Machine Learning (ICML)*, 2018.

Bao, P., She, L., McGill, M., and Tsao, D. Y. A map of object space in primate inferotemporal cortex. *Nature*, 2020.

Becker, S. Implicit learning in 3D object recognition: The importance of temporal context. *Neural Computation*, 1999.

Belharbi, S., Chatelain, C., Hérault, R., and Adam, S. Neural networks regularization through class-wise invariant representation learning. *arXiv preprint arXiv:1709.01867*, 2017.

Bengio, Y., Lee, D.-H., Bornschein, J., Mesnard, T., and Lin, Z. Towards biologically plausible deep learning. *arXiv preprint arXiv:1502.04156*, 2015.

Booth, M. and Rolls, E. T. View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cerebral Cortex*, 1998.

Cohen, T. and Welling, M. Group equivariant convolutional networks. In *International Conference on Machine Learning (ICML)*, 2016.

DiCarlo, J. J. and Cox, D. D. Untangling invariant object recognition. *Trends in Cognitive Sciences*, 2007.

Dujmović, M., Malhotra, G., and Bowers, J. What do adversarial images tell us about human vision? *bioRxiv preprint 2020.02.25.964361*, 2020.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *arXiv preprint arXiv:2004.07780*, 2020.

Hadsell, R., Chopra, S., and LeCun, Y. Dimensionality reduction by learning an invariant mapping. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2006.

Hoffer, E., Ben-Nun, T., Hubara, I., Giladi, N., Hoefler, T., and Soudry, D. Augment your batch: better training with larger batches. *arXiv preprint arXiv:1901.09335*, 2019.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. Adversarial examples are not bugs, they are features. *arXiv preprint arXiv:1905.02175*, 2019.

Isik, L., Meyers, E. M., Leibo, J. Z., and Poggio, T. The dynamics of invariant object recognition in the human visual system. *Journal of Neurophysiology*, 2013.

Jo, J. and Bengio, Y. Measuring the tendency of CNNs to learn surface statistical regularities. *arXiv preprint arXiv:1711.11561*, 2017.

Kietzmann, T. C., McClure, P., and Kriegeskorte, N. Deep neural networks in computational neuroscience. *Oxford Research Encyclopedia of Neuroscience*, 2019.

Kording, K. P., Kayser, C., Einhauser, W., and Konig, P. How are complex cell properties adapted to the statistics of natural stimuli? *Journal of Neurophysiology*, 2004.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., and Bandettini, P. A. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 2008.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. *Technical report, University of Toronto*, 2009.

Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L., and DiCarlo, J. J. Cornet: Modeling the neural mechanisms of core object recognition. *bioRxiv:408385*, 2018.

Laine, S. and Aila, T. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

Malhotra, G., Evans, B. D., and Bowers, J. S. Hiding a plane with a pixel: examining shape-bias in CNNs and the benefit of building in biological constraints. *Vision Research*, 2020.

Marblestone, A. H., Wayne, G., and Kording, K. P. Toward an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience*, 2016.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., and Fried, I. Invariant visual representation by single neurons in the human brain. *Nature*, 2005.

Richards, B. A. et al. A deep learning framework for neuroscience. *Nature Neuroscience*, 2019.

Russakovsky, O. et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 2015.

Saxe, A. M., Bansal, Y., Dapello, J., Advani, M., Kolchinsky, A., Tracey, B. D., and Cox, D. D. On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019.

Shwartz-Ziv, R. and Tishby, N. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Simard, P., Victorri, B., LeCun, Y., and Denker, J. Tangent prop-a formalism for specifying selected invariances in an adaptive network. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1992.

Sinz, F. H., Pitkow, X., Reimer, J., Bethge, M., and Tolias, A. S. Engineering a less artificial intelligence. *Neuron*, 2019.

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. Striving for simplicity: The all convolutional net. In *International Conference on Learning Representations (ICLR), arXiv:1412.6806*, 2014.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Tacchetti, A., Isik, L., and Poggio, T. A. Invariant recognition shapes neural representations of visual input. *Annual Review of Vision Science*, 2018.

Taylor, G. W., Fergus, R., LeCun, Y., and Bregler, C. Convolutional learning of spatio-temporal features. In *European Conference on Computer Vision (ECCV)*. 2010.

Tishby, N. and Zaslavsky, N. Deep learning and the information bottleneck principle. In *IEEE Information Theory Workshop (ITW)*. 2015.

Wang, H., Wu, X., Yin, P., and Xing, E. P. High frequency component helps explain the generalization of convolutional neural networks. *arXiv preprint arXiv:1905.13545*, 2019.

Wyss, R., König, P., and Verschure, P. F. J. A model of the ventral visual system based on temporal stability and local memory. *PLOS Biology*, 2006.

Zagoruyko, S. and Komodakis, N. Wide residual networks. In *British Machine Vision Conference (BMVC)*, 2016.

Zhang, R. Making convolutional networks shift-invariant again. In *International Conference on Machine Learning (ICML)*, 2019.

Chapter 7

Global visual salience of competing stimuli

Contributors

Ricardo Ramos Gameiro designed the first prototype of the experiments, collected the eye tracking data and contributed to the original draft of the manuscript. Alessandro Grillini contributed to the comparisons between global and local salience and to the corresponding part of the original draft. Peter König contributed to the conceptualisation of the project and supervised the work. Ricardo, Alessandro and Peter reviewed and edited the manuscript submitted to the Journal of Vision.

Outreach

This chapter extends the following publications:

- *Global visual salience of competing stimuli.* **Alex Hernández-García**, Ricardo Ramos Gameiro, Alessandro Grillini, Peter König. PsyArXiv preprint PsyArXiv:z7qp5 & Journal of Vision (accepted), 2019.

Visual attention is a highly complex mechanism that facilitates our understanding and navigation of the world around us, by enabling the coherent processing of the large amount of information that enters our eyes. Therefore, a fundamental component of vision and hence cognition is the guidance of eye movements (Liversedge & Findlay, 2000; Geisler & Cormack, 2011; König et al., 2016). We constantly have to decide where to look next and which regions of interest to explore, in order to process and interpret relevant information of a scene. As a consequence, the investigation of eye movement behaviour has become a major field in many research areas (Kowler, 2011; Kaspar, 2013).

In this regard, a number of studies have shown that visual behaviour is controlled by three major mechanisms: bottom-up, top-down, and spatial biases (Desimone & Duncan, 1995; Egeth & Yantis, 1997; Kastner & Ungerleider, 2000; Corbetta & Shulman, 2002; Connor et al., 2004; Kollmorgen et al., 2010). Bottom-up factors describe features of the observed image, which attract eye fixations, involving primary contrasts, such as colour, luminance, brightness, and saturation (Itti et al., 1998; Reinagel & Zador, 1999; Baddeley & Tatler, 2006). Hence, bottom-up factors are typically based on the sensory input. In contrast, top-down factors comprise internal states of the observer (Connor et al., 2004; Kaspar, 2013). That is, eye movement behaviour is also guided by specific characteristics, such as personal motivation, specific search tasks, and emotions (Wadlinger & Isaacowitz, 2006; Einhäuser et al., 2008; Rauthmann et al., 2012; Kaspar & König, 2012). Finally, spatial properties of the image, such as the image size, and motor constraints of the visual system in the brain may affect eye movement behaviour (Ramos Gameiro et al., 2017; Ossandón et al., 2014). As a result, spatial properties and motor constraints then lead to specific bias effects, such

as the central bias in natural static images (Tatler, 2007). Thus, investigating visual behaviour necessarily implies an examination of bottom-up and top-down factors as well as spatial biases.

Based on these three mechanisms—bottom-up, top-down and spatial biases—guiding visual behaviour, Koch & Ullman (1987) first revealed a method to highlight salient points in static image scenes. Whereas this model was purely conceptual, Niebur & Koch (1996) later developed an actual implementation of salience maps. This was the first prominent proposal of topographically organised features maps that guide visual attention. Salience maps describe these topographic representations of an image scene, revealing where people will most likely look at while observing the respective scene (Itti & Koch, 2001). That is, salience maps can be interpreted as a prediction of the distribution of eye movements on images. Usually, salience maps include only bottom-up image features, predicting eye fixations on image regions with primary contrasts in colour changes, saturation, luminance or brightness among others (Itti et al., 1998). However, in their first implementation, Niebur & Koch (1996) also tried to include top-down factors to build up salience maps and thus predict where people will look at most likely in image scenes. Current state-of-the-art computational salience models are artificial neural networks pre-trained on large data sets for visual object recognition and subsequently tuned to predict fixations, as is the case of Deep Gaze II (Kümmerer et al., 2016). Such models do not rely only on bottom-up features any more, but also incorporate higher-level features learned on object recognition tasks. Still, despite the better performance on salience benchmarks, deep nets-based models seem to fail at predicting the salience driven by low-level features (Kümmerer et al., 2017).

Salience maps provide a highly accurate and robust method to predict human eye movement behaviour on static images, by relying on local features to determine which parts of an image are most salient (Niebur & Koch, 1996; Itti & Koch, 2001; Kowler, 2011). However, these methods do not provide any information about the salience of the image as whole, which may depend on both local properties and also the overall semantic and contextual information of the image. Such global salience is of great relevance when an observer is faced with two or more independent visual stimuli in one context. These combinations describe situations when several stimuli compete with each other with regard to their individual semantic content, despite being in the same overall context. Such cases appear frequently in real life, for instance when two billboards hang next to each other in a mall, when several windows are open on a computer screen, or a monitor on intensive care unit, to name a few examples. Thus, by placing two or more independent image contexts side by side, as described in the previous examples, classical salience maps may well predict eye movement behaviour within each of the individual images as a closed system, but they will most likely fail to predict visual behaviour across the whole scene involving all images. Specifically, they will fail at answering the question: which stimulus is most likely to attract the observers' visual attention?

7.1 Hypotheses and contributions

In this chapter, we present the work of a study where we postulate several hypotheses. Our primary hypothesis (H1) is that it is possible to measure and calculate the global salience of natural images. That is, the likelihood of a visual stimulus to attract the first fixation of a human observer, when it is presented in competition alongside another stimulus, can be systematically modelled. In the experiment presented here, participants were confronted with stimuli containing two individual natural images—one on the left and one on the right side of the screen—at the same time. The set of images used to build our stimuli consisted of urban, indoor and nature scenes, close-ups of human faces and scenes with people in a social context. During the observation of the image pairs, we recorded the participants' eye movements. Specifically, to characterise the global salience we were interested in the direction—left or right—of the initial saccade the participant made after the stimulus onset. For further analysis, we also collected all binary saccade decisions on all the image pairs presented to the participants. We used the behavioural data collected from the participants to train a logistic regression model that successfully predicts the location of the first fixation for a given pair of images. This allowed us to use the coefficients of

the model to characterise the likelihood of each image to attract the first fixation, relative to the other images in the set. In general, images that were fixated more often are ranked higher than other images. Hence, we computed a unique *attraction score* for each image that we denote *global salience*, which depends on the individual contextual information of the image as a whole.

We also analysed the local salience properties of the individual images and compared it to the global salience. We hereby claimed that the global salience cannot be explained by the feature-driven salience maps. Formally, we hypothesise that (H2): Natural images have a specific global salience, independent of their local salience properties, that characterises their likelihood to attract the first fixation of human observers, when presented alongside another competing stimulus. A larger global salience leads to a higher attraction of initial eye movements.

In order to properly calculate the global salience, we accounted for general effects of visual behaviour in stimuli with two paired images. Previous studies have shown that humans tend to exhibit a left bias in scanning visual stimuli. Barton et al. (2006) showed that subjects looking at faces fixated longer the eye on their left side, even if the faces were inverted, and the effect was later confirmed and extended to dogs and monkeys (Guo et al., 2009). For an extensive review about spatial biases see the work by Ossandón et al. (2014), where the authors presented evidence of a marked initial left bias in right-handers, but not in left-handers, regardless of their habitual reading direction. In sum, there is a large body of evidence of lateral asymmetry in viewing behaviour, although the specific sources are yet to be fully confirmed. With respect to our study, we hypothesise that (H3): Presenting images in horizontal pairs leads to a general spatial bias in favour of the image on the left side.

In addition to the general left bias, in half of the trials of the experimental sessions, one of the images had been already seen by the participant in a previous trial, while the other was new. The participants also had to indicate which of the images was new or old. Thus, we also addressed the questions of whether the familiarity with one of the images or the task have any effect in the visual behaviour and thus in the global salience of the images. Do images that show the task-relevant scene attract more initial saccades? Likewise, are novel images more likely to attract the first fixation? This challenge sheds some light on central-peripheral interaction in visual processing. Guo (2007), for instance, showed that during face-processing, humans indeed rely on top-down information in scanning images. However, Açık et al. (2010) proposed that young adults usually rely on bottom-up rather than top-down information during visual search. In this regard, we thus hypothesise that (H4): Task-relevance and familiarity of images will not lead to higher probability of being fixated first. In order to account for any spatial bias effects that could influence the global salience model, we added coefficients to the logistic regression algorithm that could potentially capture any lateral, familiarity and bias effects. This not only makes the model more accurate, but allows us to analyse the influence of these effects. Furthermore, the location of the images in the experiments was randomised across trials and participants.

Finally, in order to better understand the properties of the global salience of competing stimuli, we also analysed the exploration time of each image. In this regard, we hypothesise the following (H5): Images with larger global salience will be explored longer than images with low global salience.

7.2 Methods: experimental setup

The present study was conducted in the Neurobiopsychology lab at the Institute of Cognitive Science of the University of Osnabrück, Germany. The experimental methods were approved by the Ethical Committee of the University of Osnabrück, Germany, and performed in accordance with the guidelines of the German Psychological Society. All participants gave written consent to participate in this study.

7.2.1 Participants

Forty-nine healthy participants (33 females, mean age = 22.39 years, *SD* = 3.63) with normal or corrected-to-normal vision took part in this study. All participants were instructed to freely observe the stimuli on the screen. In part of the measurements, they had to indicate after the trial the old or new image of a pair as further described below.

7.2.2 Apparatuses

We presented the stimuli on a 32″ widescreen Samsung monitor (Apple, California, USA) with a native resolution of 3840 $\times$ 2160 pixels. For eye movement recordings, we used a stationary Eye Link 1000 eye tracker (SR Research Ltd.) providing binocular recordings with one head camera and two eye cameras with sampling rate of 500 Hz.

Participants were seated in a darkened room at a distance of 80 cm from the monitor, resulting in 80.4 pixels per visual degree in the centre of the monitor. We did not fixate the participant's head with a headrest but verbally instructed the participants not to make head movements during the experiment. This facilitated comfortable conditions for the participants. However, eye tracker constantly recorded four edge markers on the screen with the head camera, in order to correct for small head movements. This guaranteed stable gaze recordings based on eye movements, independent of residual involuntary head movements.

The eye tracker measured binocular eye movements. For calibration of the eye tracking camera, each participant had to fixate on 13 black circles (size 0.5°) that appeared consecutively at different screen locations. The calibration was validated afterwards by calculating the drift error for each point. The calibration was repeated until the system reached an average accuracy of $< 0.5°$ for both eyes of the participant.

7.2.3 Stimuli

The images set consisted of 200 images, of which 197 were natural photographs and 3 were randomly generated pink noise images. Altogether, the stimulus set was divided into 6 categories, according to the image content: human faces, urban scenes, natural landscapes, indoor scenes, social activities and pink noise. All the photographs were obtained from either the internal image database of the Neurobiopsychology laboratory at the University of Osnabrück, Germany or the NimStim database. Each image was scaled to a resolution of 1800 $\times$ 1440 pixels. Some examples are shown in Figure 7.1a.

Each trial consisted of one stimulus with a resolution of 3840 $\times$ 2160 pixels, matching the full-size screen resolution of the display monitor (32″ diagonal; 47.8° $\times$ 26.9°). Within each presented stimulus, two images were randomly paired that is, one image was shown on the left screen side and the other image on the right screen side. Between both images, each stimulus contained a central gap of 240 pixels, as illustrated by Figure 7.1b. The background area of the stimuli was set to middle grey.

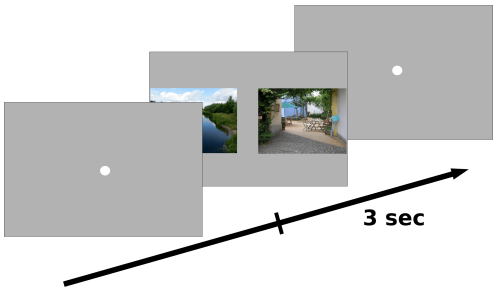
7.2.4 Procedure

The experiment consisted of 200 trials divided into four blocks, at the beginning of which the eye-tracking system was re-calibrated. The blocks were designed such that each had a different combination of task and image novelty:

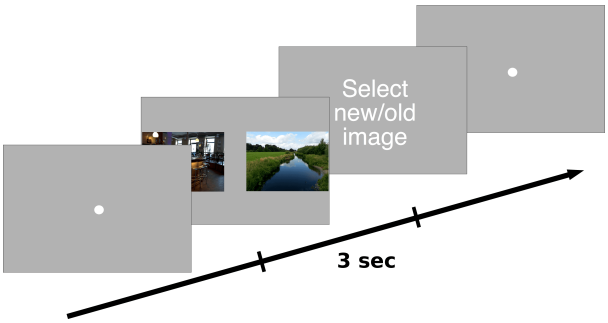
- **Block 1** consisted of 25 trials formed by 50 distinct, novel images (new/new). This block was task-free, that is participants were guided to freely observe the stimuli (Figure 7.1c).
- **Block 2** consisted of 75 trials, each formed by one new image and one of the previously seen images. (new/old or old/new). In this block, the participants were guided to freely observe

(a) Four images from the set

(b) The stimulus at one trial



(c) Procedure in blocks 1 and 4, where no task was given



(d) Procedure in blocks 2 and 3, where participants had to perform a simple task after the stimulus offset

Figure 7.1: Experimental setup

the stimuli and, additionally, they were asked to indicate the *new* image of the pair after the stimulus offset (Figure 7.1d).

- **Block 3** consisted of 75 trials, each formed by one new image and one of the previously seen images. (new/old or old/new). In this block, the participants were asked to indicate the *old* image of the pair.
- **Block 4** consisted of 25 trials formed by 50 previously seen images (old/old). Like block 1, this block was also task-free.

The decision in blocks 2 and 3 was indicated by either pressing the left—task-relevant image is on the left side—or right—task relevant-image is on the right side—arrow button on a computer keyboard.

The image pairs were formed by randomly sampling from the set of 200 images, but some constraints were set in order to satisfy the characteristics of each block and keep a balance in the number of times each image was seen by the participant. The sampling process was as follows: In block 1, 50 images were randomly sampled to form the 25 pairs. In blocks 2 and 3, in order to construct the new/old and old/new pairs, the new image was randomly sampled from the set of remaining unseen images and the old image was randomly sampled of previously seen images, with two additional constraints: it must have been chosen only one time before and not in the previous 5 trials. Finally, in block 4, a set of exactly 50 images which had been shown only once remained. These were used to randomly sample the remaining 25 trials. In all blocks, after sampling the two images, the left/right configuration was also randomly chosen with probability 0.5.

The sampling process was different for each participant, that is they saw different sets of pairs

from the 40,000 different pairs and in different order. This aimed at reducing the predictability of the process while satisfying the experimental constraints. Overall, we collected data from 9,800 pairs, some of which might have been repeated across participants. However, note that each participant saw each image exactly twice, therefore the frequency of presentation of the images was balanced across the whole experiment. As we will see in the following section, the amount of data was enough to fit the computational model.

In all cases, the presentation time for each stimulus was 3 seconds and it was always preceded by a blank, grey screen with a white, central fixation dot. The stimulus was displayed only after the participant fixated the central dot.

The majority of our analyses focused on the first fixation. As a pre-processing stage, we discarded the fixations 1) due to anticipatory saccades, 2) shorter than 50 ms or longer than $\mu_{dur} + 2\sigma_{dur}$ ms, where $\mu_{dur} = 198$ ms and $\sigma_{dur} = 90$ ms are the mean and standard deviation of all fixation durations, respectively, and 3) located outside any of the two images. The discarded fixations were less than 4 % of the total.

7.3 Methods: computation of global salience

In order to characterise the global salience of competing stimuli, we trained a logistic regression model with the behavioural data from the eye-tracking experiments. Provided that the model can accurately predict the location of the first fixation—left or right—the coefficients for each image will represent the likelihood of the image to attract the first fixation and this, in turn, can then be interpreted as the global image salience. The intuition is that images that are more often the target of the first fixation after the stimulus onset have a higher global salience, and vice versa.

7.3.1 Logistic regression for pairwise estimates

Typically, logistic regression is used in binary classification problems, as is this case where the initial fixation after stimulus onset can land either on the left ($y = -1$) or on the right ($y = 1$) image. The classifier simply estimates a probability $h_w(\mathbf{x})$ for the binary event on the linear hypothesis $\mathbf{w}^T\mathbf{x}$ by applying a logistic function:

$$h_w(\mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}^T\mathbf{x}}} = \frac{e^{\mathbf{w}^T\mathbf{x}}}{1 + e^{\mathbf{w}^T\mathbf{x}}} \tag{7.1}$$

where $\mathbf{x}$ is a vector that represents the independent or explanatory variables (features) and $\mathbf{w}$ the coefficients to be learned. Thus, the likelihood of the binary outcome given the data is the following:

$$P(y|\mathbf{x}) = \begin{cases} h_w(\mathbf{x}) & \text{if } y = 1 \\ 1 - h_w(\mathbf{x}) & \text{if } y = -1 \end{cases} = \frac{e^{y\mathbf{w}^T\mathbf{x}}}{1 + e^{y\mathbf{w}^T\mathbf{x}}}$$

The coefficients are then optimised by minimising the negative log-likelihood $-log(P(y|\mathbf{x}))$ through gradient descent. Typically, a regularisation penalty is added on the coefficients, controlled by the parameter C—inverse of the regularisation strength. In our case, we applied L_2 regularisation and therefore the algorithm solves the following optimisation problem, given a set of N training data points (trials):

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N} log(1 + e^{-y_i\mathbf{w}^T\mathbf{x}_i}) \tag{7.2}$$

The optimisation problem was solved through the *LIBLINEAR* algorithm (Fan et al., 2008), available in the `scikit-learn` Python toolbox.

In our particular case, for every trial i—stimulus pair seen by a participant—each feature x_{ij} corresponded to one image j and only two images were shown at each trial. Therefore, we were interested in modelling the probability that one image u receives the first fixation when presented next to another image v; hence $p(u > v)$. This simplifies the standard logistic regression model to a special case for pairwise probability estimates, known as the Bradley-Terry-Luce model (Bradley & Terry, 1952; Luce, 2005), where the probability h_w is the following:

$$h_w(u, v) = p(u > v) = \frac{e^{w_u}}{e^{w_u} + e^{w_v}} = \frac{e^{w_u - w_v}}{1 + e^{w_u - w_v}} \tag{7.3}$$

where w_u and w_v are the coefficients of image u and v. This is a special case of the function in Equation 7.1, where all the elements in the feature vector $\mathbf{x}$ are zero except for the two paired features x_u and x_v, which are set to 1 and -1 respectively. Note that in the Bradley-Terry-Luce model the coefficients still refer to the whole set of features and therefore are described by an M-dimensional vector $\mathbf{w} = \{w_1, w_2, ..., w_M\}$, where in our case $M = 200$, the total number of images in the set. After training the model, each learned coefficient w_j will be related to the average likelihood of image j of receiving the first fixation when presented next to other images from the set. As stated above, we interpret these coefficients $\mathbf{w}$ as a measure of the global image salience.

In order to estimate the coefficients $\mathbf{w}$, the logistic regression model was trained on the data set arranged into a design matrix X of the following form:

$$X = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \dots & x_M^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \dots & x_M^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(N)} & x_2^{(N)} & \dots & x_M^{(N)} \end{bmatrix} \tag{7.4}$$

where each row represents one measured data point: one trial where one participant was presented a pair of images u and v—the total number of trials was in our case $N = 49$ participants $\times$ 200 trials per participant $= 9800$—and where the columns represent the values of the different features (images) that were tested ($M = 200$). According to Equation 7.3, if image u is presented on the right and image v is presented on the left at trial i, then $x_u^{(i)} = 1$, $x_v^{(i)} = -1$ and $x_j^{(i)} = 0, \forall\, j \neq u, v$. Finally, the outcome of each trial is given as a vector $\mathbf{y}$:

$$\mathbf{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)} \end{bmatrix}$$

such that $y^{(i)} = 1$ if the right image was fixated first, and $y^{(i)} = -1$ if the left image was fixated first at trial i.

7.3.2 Task, familiarity and lateral bias

Not only were we interested in modelling the likelihood of every image of receiving the first fixation, but also the contribution of other aspects of the experiment, namely the effect of having to perform a small task when observing the pair of images and the familiarity with one of the two images. More specifically, we were interested in answering the following questions: Do light task demands, such as having to determine which image is new or old, influence the direction of the first saccade? And, are unseen stimuli more likely to receive the initial saccade than previously observed stimuli when presented together, or vice versa?

We addressed these questions by adding new features to the model that capture these characteristics of the experimental setup. These features were assigned coefficients that, after training, will indicate the magnitude of the contributions of the effects. In particular, we added the following features columns to every row i of the design matrix:

- $t^{(i)}$: 1 if the target of the task (select new/old image) was on the right at trial i, -1 image if on the left, 0 if no task.
- $f^{(i)}$: 1 if at trial i, the image on the right had been already shown at a previous trial (familiar), while the image on the left was still unseen; -1 if the familiar image was on the left; 0 if both images were new or familiar.

Not only did these new features enable new elements for the analysis, but also added more representational power to the model, which could potentially learn better coefficients to describe the global salience of each image. In this line, we added one more feature to the model to capture one important aspect of visual exploration: the lateral bias. Although a single intercept term in the argument of the logistic function ($\mathbf{w}^T\mathbf{x} + b$) would capture most of the lateral bias, since the outcome $\mathbf{y}$ describes exactly the lateral direction, left or right, of the first saccade, we instead added subject-specific features to model the fact that the trials were generated by different subjects with an individual lateral bias. This was done by adding $K = 49$ (number of participants) features $s_k^{(i)}$, with value 1 if the trial i was performed by subject k and 0 otherwise. Altogether, the final design matrix X' extends the design matrix X defined in Equation 7.4 as follows:

$$X' = \left[\begin{array}{ccc|c|c|ccc} x_1^{(1)} & \dots & x_M^{(1)} & t^{(1)} & f^{(1)} & s_1^{(1)} & \dots & s_K^{(1)} \\ x_1^{(2)} & \dots & x_M^{(2)} & t^{(2)} & f^{(2)} & s_1^{(2)} & \dots & s_K^{(2)} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1^{(N)} & \dots & x_M^{(N)} & t^{(N)} & f^{(N)} & s_1^{(N)} & \dots & s_K^{(N)} \end{array}\right] \tag{7.5}$$

Note that the leftmost block of X' is identical to X (defined in Equation 7.4). While the shape of X is 9800×200, X' is a 9800×251 matrix, since $200 + 1 + 1 + 49 = 251$.

7.3.3 Validation and evaluation of the model

In order to ensure the successful training of the model, we carried out a 5-fold cross-validation of the regularisation parameter C of the model, described in Equation 7.2. That is, we split our data set into 5 different folds of 39 subjects for training and 10 for validation—7,800 and 2,000 trials, respectively—and evaluated the performance with 10 different values of C, according to the following search space:

$$C = 10^p \quad \text{with } p = -3 + \frac{2}{3}(n-1) \quad \text{and } n = 1, ..., 10$$

The value that provided the best average performance across the folds was selected.

In order to reliably assess the model performance while taking the most out of the data set, we embedded the cross-validated model into a *leave-2-participants-out* cross-evaluation. That is, we constructed 25 different folds of data, each with the trials of 23 participants for training and of 2 participants for evaluation. We report here the average performance across the 25 test and train partitions together with the standard deviation (within brackets). In particular, in Table 7.1 we include the area under the curve (AUC), the Tjur[1] coefficient of discrimination R^2 and the accuracy. For the sake of an easier interpretation, we include the theoretical baseline values of the AUC and R^2, and the empirical baseline accuracy on our test partitions.

[1]While there is no consensus about the best metric for the evaluation of logistic regression, the coefficient of discrimination R^2 proposed by Tjur (2009) has been widely adopted recently, as it is more intuitive than other definitions of coefficients of determination and still asymptotically related to them.

	AUC	Tjur R^2	Accuracy
Test	0.8884 (0.0180)	0.4287 (0.0460)	81.36 % (0.32)
Train	0.8865 (0.0040)	0.4240 (0.0214)	81.99 % (1.52)
Random baseline	0.5	0.0	60.70 % (2.32)

Table 7.1: Test, train and baseline performance of the logistic regression model. Values within brackets indicate the standard deviation across the folds.

The results in Table 7.1 show that the logistic regression model successfully learned the behavioural patterns from the experimental data and hence accurately predicted the direction of the first saccade, with very low overfitting, since train and test performance were very similar and have low variance. As a conclusion, this implies that the learned coefficients can be meaningfully used for further analysis, as will be presented in Section 7.5.

7.4 Methods: salience maps of competing stimuli

In order to test whether the global salience is independent from the lower-level, salience properties of the stimuli (H2), we also computed salience maps both of each individual image and of each full stimulus shown at each trial, that is the pair of images with grey background, as shown in Figure 7.1b. For the computation of the salience maps we used the Graph-Based Visual Salience algorithm (GBVS) (Harel et al., 2007), which is a computational salience model that makes use of well-defined low-level features.

Moreover, we also analysed the connection between global salience and a less restricted salience model, Deep Gaze II (Kümmerer et al., 2016), whose features include higher level cues, since it is a deep neural network model pre-trained for large scale, image object recognition tasks, with additional features optimised for salience prediction.

In order to compare the salience maps with the behavioural data from the observation of competing stimuli, as well as with our derived global salience, we performed the following tests:

7.4.1 Predictivity of salience maps for the first fixation

In this case, our aim was to evaluate the performance of salience maps in predicting the landing location of the first fixation when two competing stimuli are presented. To do so, we computed the Kullback-Leibler Divergence between the first fixation distribution $F_j(b)$ and the salience distribution $S_j(b)$ for every image j in the set of 200 images:

$$D_{KL}(F_j||S_j) = \sum_{b=1}^{B} F_j(b) \log(\frac{F_j(b)}{S_j(b) + \epsilon} + \epsilon) \tag{7.6}$$

where is ϵ is a small constant to ensure numerical stability and b refers to B bins of one 1×1 degrees of visual field angle.

The first fixation distribution, $F_j(b)$, is the probability density distribution of all the first fixations made by all observers on each image j. To compute $F_j(b)$, we divided every image into sections of one squared degree of visual field angle and counted the number of first fixations made by all participants on each bin to obtain a histogram. Then, the histogram was smoothed using a Gaussian kernel with a size of one degree of visual field angle and normalised such that it became a probability distribution. The salience distribution, $S_j(b)$, is the smoothed and normalised (likewise) salience map—computed with GBVS or Deep Gaze II—of each individual image j.

Hence, according to the definition in Equation 7.6, a low $D_{KL}(F_j||S_j)$ would imply a good match between the location of the first fixations and the salience map of image j.

7.4.2 Comparison between global and local salience

In order to compare the local salience maps and the global salience scores learned by the computational model presented in Section 7.3, we analysed the GBVS and Deep Gaze salience maps of both the individual images and the whole stimuli, in relation to the global salience scores.

Individual images

First, we compared the Kullback-Leibler Divergence between the first fixations distribution and the salience maps of the individual images, as computed in Equation 7.6, and the global salience scores, that is the coefficients learned by the optimisation defined in Equation 7.2. This aimed at analysing whether, for instance, images whose local salience properties indeed drove the location of the first fixation have a higher global salience score, and vice versa.

Trials

Second, we looked at the properties of the salience map of the final stimulus seen by the participants at each trial, that is the paired competing images with a grey background (see Figure 7.1b). As a metric of the contribution of each image to the salience map, for each trial i we computed the relative salience mass M of each image, left and right:

$$M_i^L = \int_{x \in X_L} S_i(x) \quad M_i^R = \int_{x \in X_R} S_i(x)$$

where $S_i(x)$ is the normalised salience map of the whole stimulus presented at trial i and X_L and X_R are the stimulus locations corresponding to the left and right images, respectively. A significant positive correlation between $\Delta_M^{(i)} = M_i^L - M_i^R$ and the difference between the global salience scores of the images on the left and right, $\Delta_{GS}^{(i)} = w_L^{(i)} - w_R^{(i)}$, would indicate that the local salience properties can partly explain the direction of the first fixation.

7.5 Results

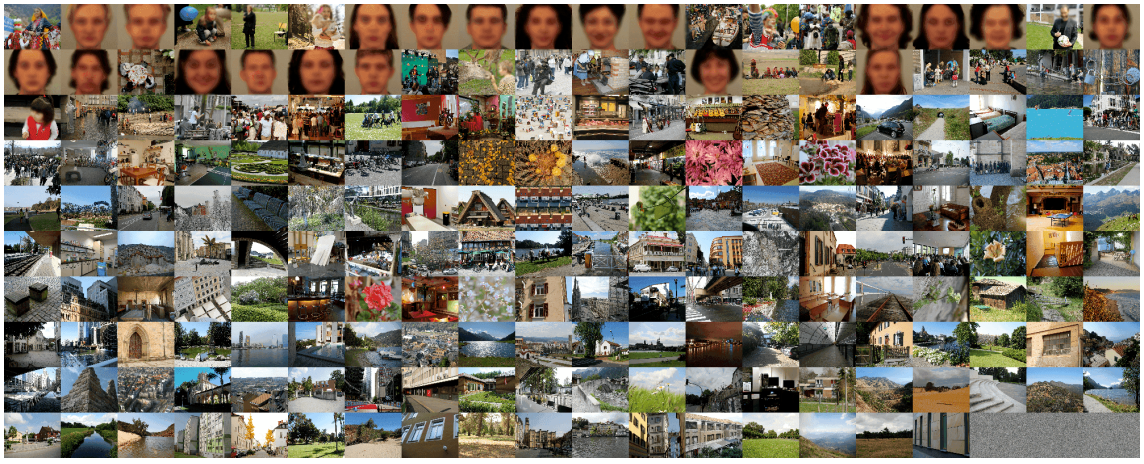
In this section, we present the main results of our analyses and discuss the validity of the hypotheses presented in Section 7.1. Each of the sub-sections focuses on one of the five hypotheses, in the natural order. All the scatter plots that show the relationship between two variables include the value of the Pearson correlation, as well as the line fit by a linear regression model, with 95 % confidence intervals estimated using bootstrap with 1,000 resamples.

7.5.1 Global visual salience

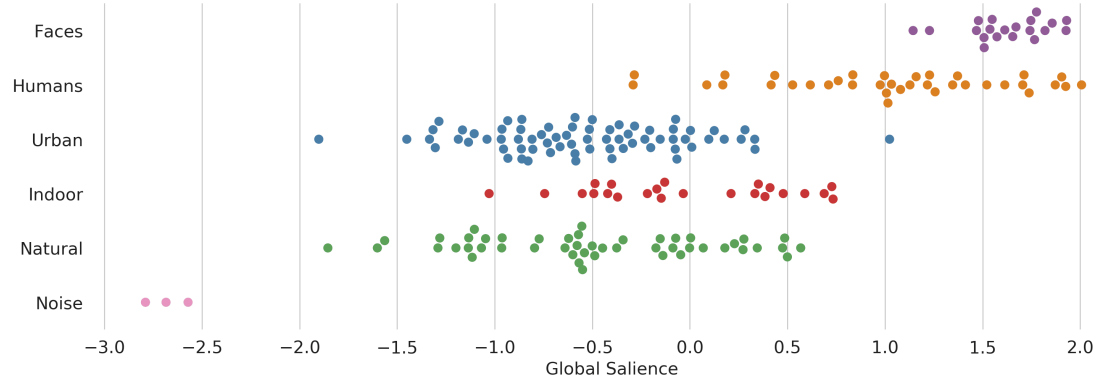
In our first hypothesis (H1), we stated that images can be ranked according to a specific global salience that leads to the attraction of initial eye fixations. In order to quantify the global salience of individual images, we have presented in Section 7.3 a computational model that successfully predicts the direction of the first fixation from the behavioural data, as validated by the results in Section 7.3.3, and thus we can analyse the coefficients of the model as indicators of the global salience of each image in the data set.

Importantly, the fact that the first fixation direction of the participants when exploring such competitive stimuli can be predicted by a computational model means that their behaviour was

not random, but followed certain patterns. In order to shed some light on the nature of these patterns, in Figure 7.2a we show the complete set of stimuli ranked according the global salience score learned by our model and in Figure 7.2b the value of the global salience scores of each image, highlighting the differences between the image categories.



(a) Experimental stimuli, ranked according to the learned global salience. The stimulus with the highest global salience score is on the top-left corner and the rest are sorted with the x-axis changing fastest (row-major order). Faces have been blurred to preserve the identity.



(b) Global salience score of each stimulus and image categories.

Figure 7.2: Global salience scores of the experimental stimuli

Figure 7.2 shows that there exists a clear, general tendency to first fixate on the images that contain either close-up faces or scenes with humans, even though the first fixations may occur, on average, as early as after 242 ms ($\sigma_{SD} = 66$ ms) from the stimulus onset. These two categories, faces and humans, were assigned the highest global salience scores. Then, urban, indoor and natural landscapes obtained significantly lower scores, with no big differences among the three categories. Finally, the three pink-noise images were assigned very low scores, which serves as sanity-check of our proposed method.

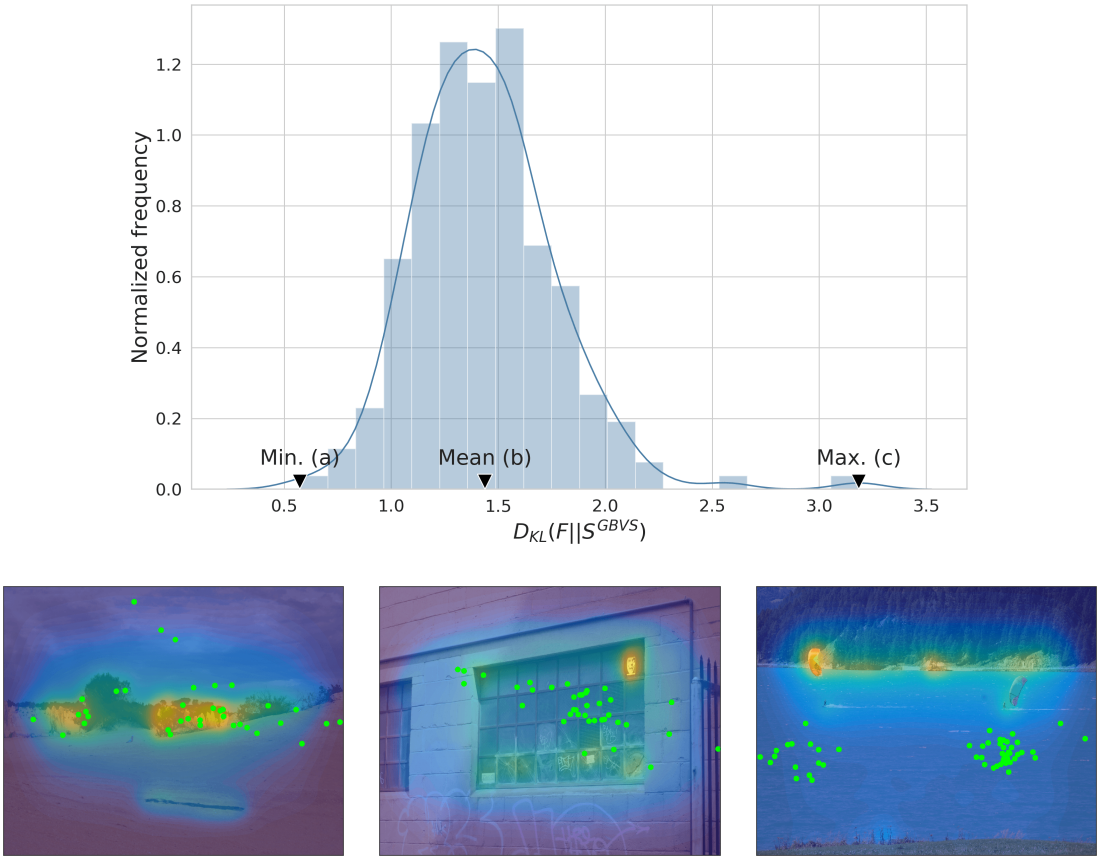
7.5.2 Global vs. local salience

A reasonable question in view of the results presented in Figure 7.2 is whether the global salience scores—and the ranking of the stimuli that arises from the scores—is a unique measure that assesses the initial visual behaviour when facing competing stimuli, or whether this behaviour and thus our proposed global salience can be explained by the low-level properties of the respective stimuli.

In our second hypothesis (H2) we stated, instead, that the global salience is independent from

the low-level local salience properties. So as to test this, we performed several tests, described in Section 7.4.

In Figure 7.3 we plot the distribution of the Kullback-Leibler Divergence between the first fixations maps and the GBVS local salience maps of the individual images (see Equation 7.6). The mean of the distribution is significantly non-zero (two-tail t-test p-value $< .001$, $\mu_{KLD} = 1.44$, $\sigma_{KLD} = 0.33$), which means that there is a significant loss of information when using a local salience map to predict the landing locations of the first fixations on a given image (Riche et al., 2013). In order to illustrate the mismatch, in Figure 7.3 we display three example images with the overlaid salience maps and the location of all the first fixations that landed on them. When the KLD value is minimum (a), the salience maps can approximate the fixations, although this happened rarely. Already with KLD values around the mean, the performance of a salience map in predicting the landing location of fixations is rather mediocre (b) and deteriorates further as the KLD increases (c).

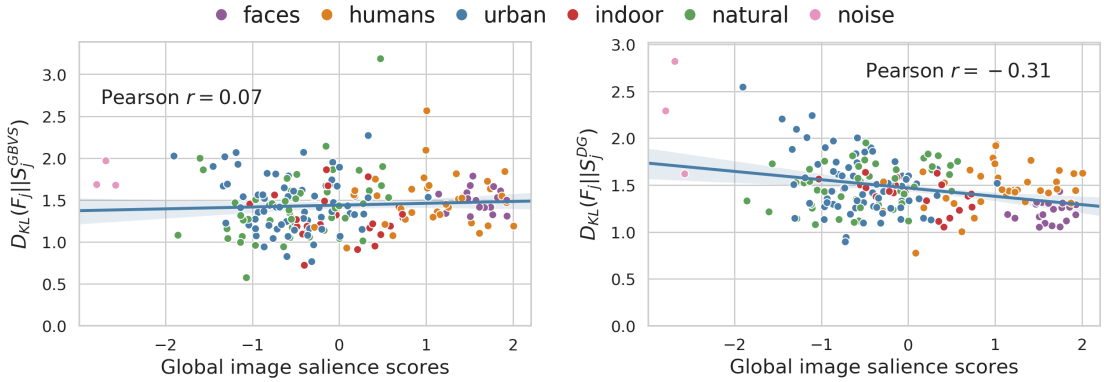


(a) $D_{KL}(F||S) = 0.57$ (b) $D_{KL}(F||S) = 1.44$ (c) $D_{KL}(F||S) = 3.19$

Figure 7.3: Top row: distribution of the Kullback-Leibler Divergence between the first fixations map and the GBVS local salience maps. Bottom row: images with the minimum, closest to the mean and maximum KLD, with their overlaid salience map and the location of the first fixations.

Perhaps not surprisingly, in view of the poor match between the salience maps and the first fixation maps, Figure 7.4a shows that the Kullback-Leibler Divergence between them does not correlate with the global salience scores. This means that the images which attract the first fixations towards salient regions (low KLD) do not tend to have high global salience scores neither vice versa.

Finally, we analyse in Figure 7.5 whether the direction of the first fixation when looking at competing stimuli, as modelled by our proposed global salience scores, can be explained by the difference in the low-level salience properties of the competing stimuli, as measured by the GBVS



(a) GBVS

(b) Deep Gaze

Figure 7.4: Comparison between the global salience scores the KLD between the first fixation distribution and the salience maps from the computational models

salience mass of each image (see Section 7.4.2). Also in this case we found no significant correlation.

The noisy images included in the stimulus set serve once more as a validation of the expected results. When one of the images (left or right) was pink noise the difference in GBVS salience mass was either very high or very low, as is the difference in global salience scores. In this case, both metrics do correlate, but as shown by the central scatter plot of Figure 7.5, the feature-driven (GBVS) salience mass cannot explain the global salience scores learned by the model.

In order to better understand what drives the direction of the first fixation when faced with competing stimuli, we also compared our proposed global salience with properties of Deep Gaze II salience maps. As presented in Section 7.4, unlike GBVS, Deep Gaze does make use of higher-level information of the images to predict the salience maps, since it is a neural network pre-trained on image object recognition tasks. This allows it to model salience driven by faces or objects (Kümmerer et al., 2017) and it becomes an interesting model to which compare our global salience model, since we have seen in Section 7.5.1 that images containing faces and humans tend to get a higher global salience score.

In general, we observe that unlike GBVS, measures derived from Deep Gaze salience maps exhibit a non-zero, yet moderate correlation with our proposed global salience. For instance, Figure 7.4b shows a slight negative correlation between global salience scores and the KLD between first fixation distributions and Deep Gaze salience maps. However, looking at the distribution of the Kullback-Leibler divergence in Figure 7.6, we see that the salience maps are also far from matching the location of the first fixations on the images. Finally, we also observed (see Figure 7.7 a non-zero correlation between the difference of global salience scores between the left and the right image, and the difference in salience mass computed with Deep Gaze.

Taken together, we can conclude that our proposed computational model provided a robust method to rank images according to a unique global image salience that is independent of the low-level local salience properties of the stimuli, and we observed a non-zero, yet moderate correlation with a computational salience model that incorporates higher-level cues.

7.5.3 Lateral bias

Our third hypothesis (H3) stated that a general spatial bias leads to a higher likelihood to first fixate on the left rather than the right image. We thus calculated the number of first saccades that landed onto the left and the right image for each block separately (Figure 7.8). A 4×2 (block: 1, 2, 3, $4 \times$ image side: left, right) repeated measures ANOVA (Greenhouse-Geisser corrected) revealed a general spatial bias of the initial saccade towards the left image as indicated by a significant main

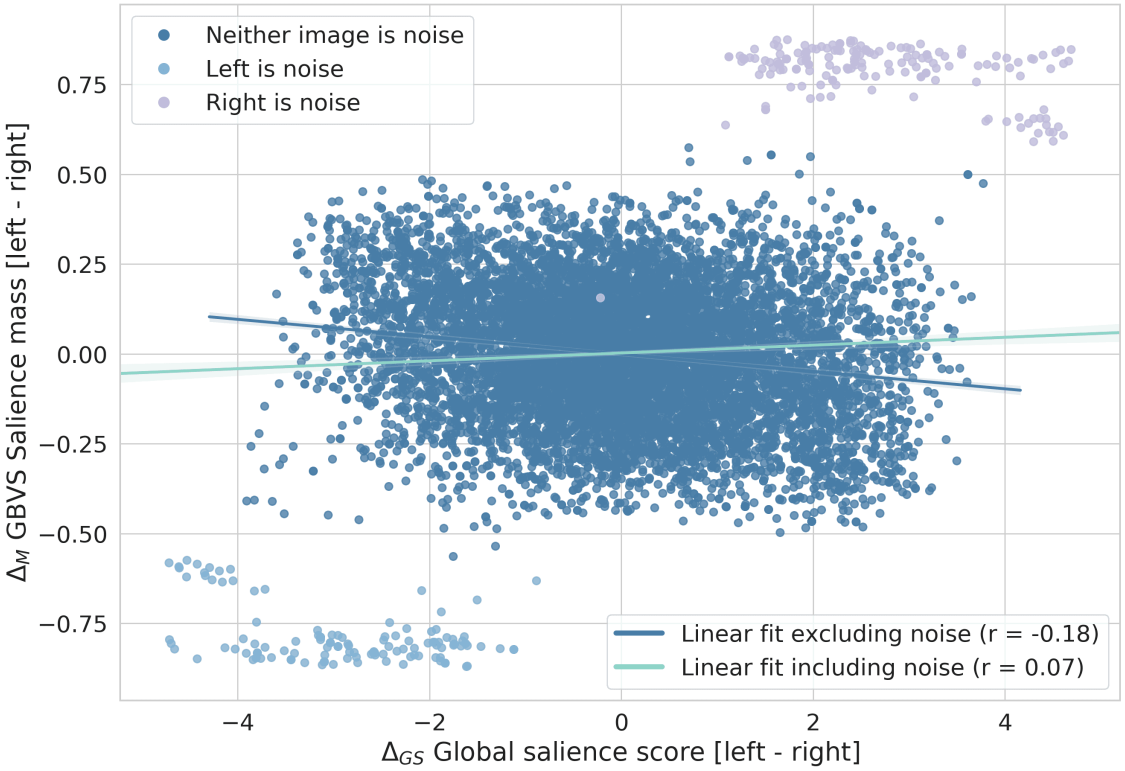


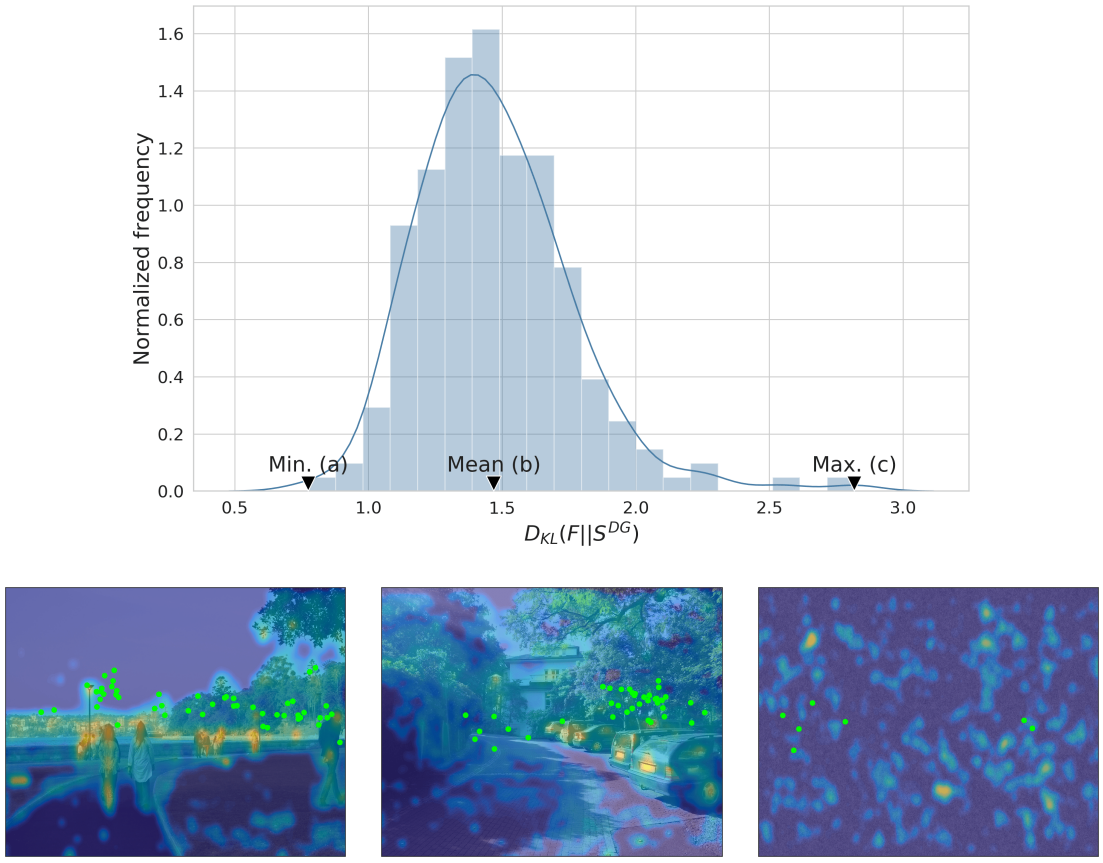
Figure 7.5: Correlation between the GBVS image salience mass and the global salience scores.

effect according to the image side [$F(1, 48) = 30.833$; $p < .001$; $\eta_p^2 = .391$]. No further effects were found [all $F \leq 2.594$; all $p \geq .074$, all $\eta_p^2 \leq .051$], showing that the left bias was present in all blocks with similar extent. Thus, we can conclude that the participants generally targeted their initial saccades more on left than right sided images.

Nonetheless, the error bars in Figure 7.8 suggest a high variability of the lateral bias across subjects. In order to investigate this, we calculated the number of first saccades on the right image for each participant separately. Moreover, since our model included an individual bias term for each participant, as described in Section 7.3, we can also look at the magnitude of the coefficients learned by the model. In Figure 7.9 we plot, for each participant, the percentage of first saccades towards the right image and their corresponding lateral bias term learned by the computational model. Both metrics are highly correlated—further highlighting the validity of the model—and reveal a high variability in the lateral bias across participants. Overall, 63 % of all first fixations landed on the left image.

7.5.4 Task and familiarity

Next, we investigated the effect of the familiarity with one of the images and of the task of selecting the already seen or unseen image, which the participants had to perform in blocks 2 and 3 of the experiment, respectively. In particular, we were interested in finding out whether there is a tendency to direct the initial saccade towards the task-relevant images or towards the new images, for instance. In our fourth hypothesis (H4) we stated that our task and familiarity should have little or no influence in the initial saccade. For that purpose, we first performed a 2×2 (task: select new, select old $\times$ fixated image: new image, old image) repeated measures ANOVA analysis (Greenhouse-Geisser corrected). The results revealed no significant effects [all $F \leq 1.936$; all $p \geq .170$, all $\eta_p^2 \leq .039$] (Figure 7.10). Thus, the provided tasks did not bias the initial saccade decision to target one of the two presented images. Nevertheless, we found that participants correctly identified 91.43% of the new images in block 2 and 91.16% of the old images in block 3. Hence, the task



(a) $D_{KL}(F||S) = 0.57$ (b) $D_{KL}(F||S) = 1.44$ (c) $D_{KL}(F||S) = 3.19$

Figure 7.6: Top row: distribution of the Kullback-Leibler Divergence between the first fixations map and the Deep Gaze local salience maps. Bottom row: images with the minimum, closest to the mean and maximum KLD, with their overlaid salience map and the location of the first fixations.

performance was highly above chance (50%) and the participants were accurate in identifying the new and old images respectively.

Also in this case, the same conclusion can be extracted from the coefficients learned by the model to capture the task and familiarity effects, which are -0.04 and -0.10, respectively, that is, very small and only slightly higher for the familiarity.

Taken together, spatial properties influenced the initial saccade in favour to fixate left sided images first. Although task performance was very high, neither the task nor the familiarity with one of the images had an influence in the direction of the first fixation after stimulus onset. These results fully support our third and fourth hypotheses.

7.5.5 Total exploration of images

In our fifth hypothesis (H5), we stated that images with higher global image salience lead to a longer exploration time than images with lower global salience. We thus calculated the relative dwell time on each image, left and right, for each trial. As an initial step, similarly to the analysis of the initial saccade, we analysed the potential effect of the spatial image location as well as the task and familiarity relevance on the exploration time.

With respect to the spatial image location, a 4 $\times$ 2 (block: 1, 2, 3, 4 $\times$ image side: left, right) repeated measures ANOVA (Greenhouse-Geisser corrected) revealed a significant main effect ac-

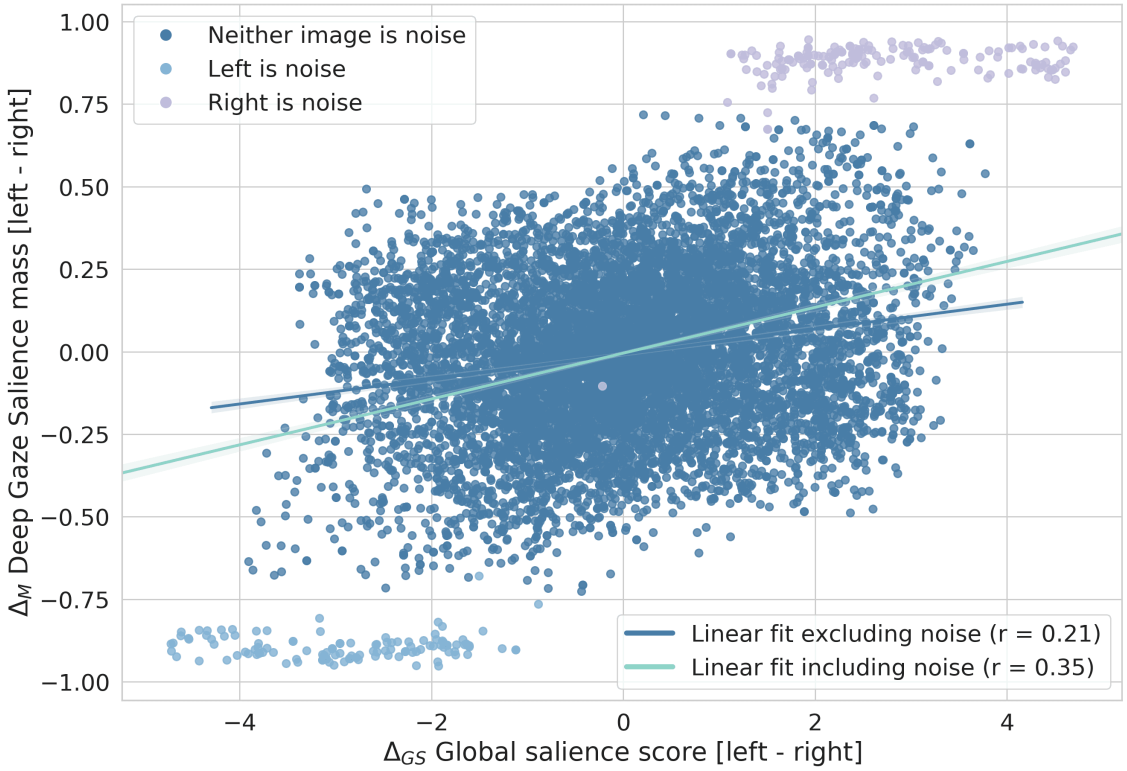


Figure 7.7: Correlation between the Deep Gaze image salience mass and the global salience scores.

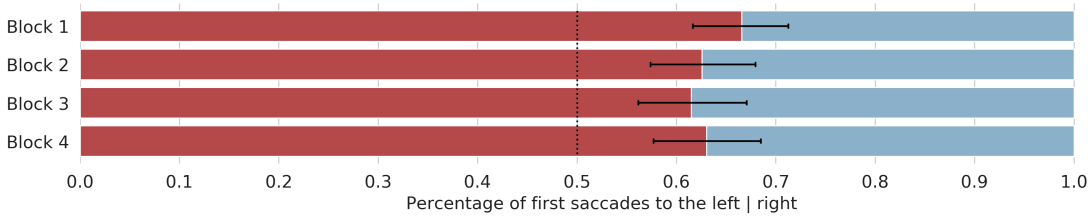


Figure 7.8: Percentage of first saccades that targeted on the left (red) and right (blue) images, at each block of the experimental session. Error bars depict the standard deviation of the mean. Note that considerably more first fixations landed on the left image, highlighting the lateral bias.

cording to the block [$F(2.368, 113.668) = 12.066$; $p < .001$; $\eta_p^2 = .201$] but no further effects [all $F \leq 2.232$; all $p \geq .109$, all $\eta_p^2 \leq .044$]. Thus, the total time of exploration did not depend on the spatial location of the images, as also shown in Figure 7.11.

With respect to the task relevance—recall: block 2 - select new image; block 3 - select old image—we calculated a 2 × 2—task: select new, select old × fixated image: new image, old image—repeated measures ANOVA (Greenhouse-Geisser corrected). The results revealed a significant main effect according to the task [$F(1, 48) = 4.298$; $p < .050$; $\eta_p^2 = .082$] and fixated image [$F(1, 48) = 64.524$; $p < .001$; $\eta_p^2 = .573$], as well as an interaction between task and fixated image [$F(1, 48) = 36.728$; $p < .001$; $\eta_p^2 = .433$]. As shown by Figure 7.12, our results showed that, in general, participants tended to spend more time exploring new instead of previously seen images. Furthermore, this effect was noticeably larger in block 2, where the task was to select the new images, than in block 3 (select old image).

Consequently, we found that the spatial location of images did not affect the total time of exploration. Instead, the task and familiarity had a considerable impact on the exploration time, revealing that new images were explored during a longer time than the counterpart.

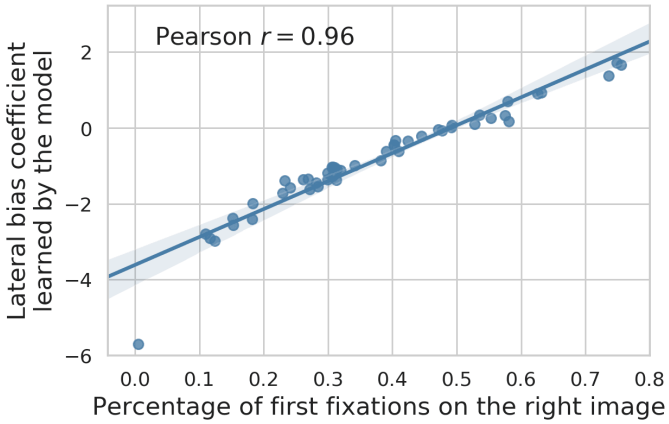


Figure 7.9: Lateral bias of each participant, as measured by the percentage of first fixations onto the right image and the lateral bias terms learned by our computational model. Both metrics are highly correlated and reveal the average left bias, but with high variability across participants.

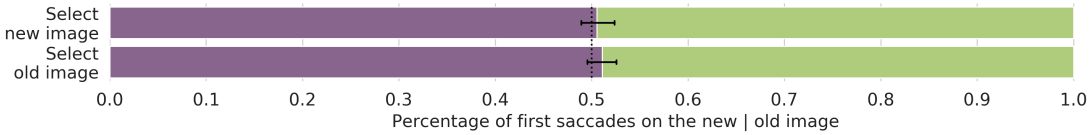


Figure 7.10: Percentage of first saccades that targeted on the new (purple) and old (green) images, at blocks 2 and 3, where participants had the task of indicating the new and old image, respectively. Error bars depict the standard deviation of the mean. No significant bias can be appreciated in this case.

For our main analysis regarding the interaction between exploration time and global image salience, we then contrasted the global salience score learned for each image with its respective dwell time averaged over all trials and subjects. The results revealed a significant positive correlation, indicating that images with larger global image salience led to a more intense exploration (Figure 7.13). Thus, global image salience describes not only a measure of which image attracts initial eye movements, but is also connected to longer exploration time, suggesting that global salience may describe the relative engagement of images.

Taken together, our results suggest that the task and familiarity—but not the spatial location of images—influenced the exploration time with respect to higher dwell times on unseen images in combination with the task to select the new image. Note, however, that regarding the effects of task our findings are restricted to the specific task assigned in our experiments, that is selecting which image is new or old. The effects of task in visual attention is an active field in visual perception and the results of multiple contributions should be taken together into consideration to draw robust conclusions. Finally, we also found that images with higher global salience correspondingly led to a larger time of exploration. These results fully support our fifth hypothesis.

7.6 Discussion

We have presented a computational model trained on the saccadic behaviour of participants freely looking at pairs of competing stimuli, which is able to learn a robust score for each image, related to its likelihood of attracting the first fixation. This fully supports our first hypothesis and we refer to this property of natural images as the global visual salience.

The computational model consists of a logistic regression classifier, trained with the behavioural data of 49 participants who were presented 200 pairs of images. In order to reliably assess

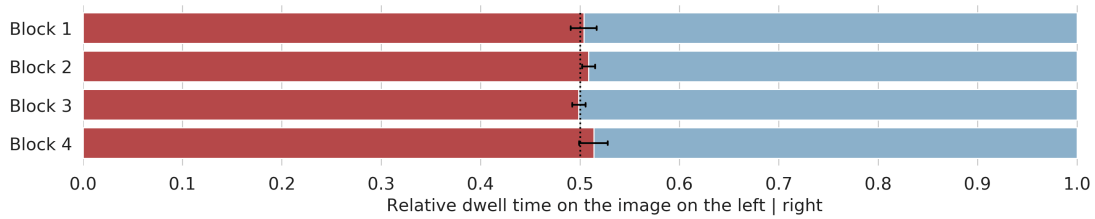


Figure 7.11: Exploration as measured by the relative dwell time on the left (red) and right (blue) images, at each block of the experimental session. Error bars depict the standard deviation of the mean.

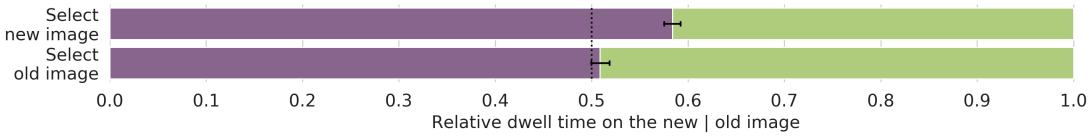


Figure 7.12: Exploration as measured by the relative dwell time on the new (purple) and old (green) images, at blocks 2 and 3, where participants had the task of indicating the new and old image, respectively. Error bars depict the standard deviation of the mean.

the performance of the model, we carried out a careful 25-fold cross-evaluation, with disjoint sets of participants for training, validating and testing. Given a pair of images from the set of 200, the model predicted the direction of the first saccade with 82 % accuracy and 0.88 area under the ROC curve.

Throughout this chapter, we have analysed the general lateral bias towards the left image (H2), as well as other possible influences such as the familiarity with one of the images and the effect of a simple task (H3). Moreover, we have analysed the relationship of our proposed global salience with the local salience properties of the individual images (H4). Finally, we have also studied the total exploration time of each image in the eye-tracking experiment and compared it to the global salience, which is based upon the first fixation (H5).

Regarding the lateral bias, we found that participants tended to look more frequently towards the image on the left. Such left bias is typical in visual behaviour and has been found in many previous studies (Barton et al., 2006; Guo et al., 2009; Calen Walshe & Nuthmann, 2014; Ossandón et al., 2014). However, most of these studies presented only single images per stimulus. In this regard, it has been argued that cultural factors of the Western population who mostly take part in the research experiments may lead to a semantic processing of natural visual stimuli similar to the reading direction, that is from left to right (Spalek & Hammad, 2005; Zaeinab et al., 2016).

In our study, about 63 % of the first fixations landed on the left image. However, we also observed a high variability across participants, successfully captured by our computational model. In contrast, we showed that the given task in certain trials did not influence initial saccade behaviour. Participants equally distributed the target location of saccades on the presented images, regardless of familiarity and task relevance. Consequently, the spatial location of an image affected saccade behaviour, whereas the task as well as familiarity had no influence.

Importantly, we found that that global salience, that is the likelihood of an image attracting the first fixation when presented next another competing image, is independent of the low-level local salience properties of the respective images. The location of the first fixations made by the participants in the study did not correlate with the GBVS salience maps of the images and the saccadic choice—left or right—was neither explained by the GBVS salience mass difference. Hence, our results provide some new insights in the understanding of visual perception of natural images, showing that the global salience of an image is rather affected by the semantics of the content. For instance, images involving socially relevant content such as humans or faces led to higher global salience than images containing purely indoor, urban or natural scenes.

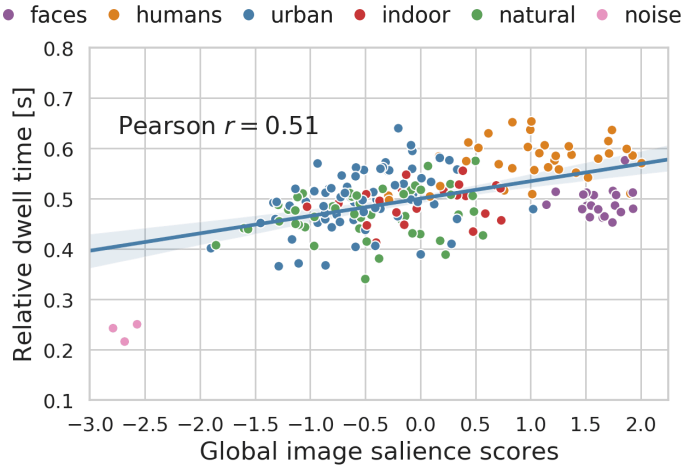


Figure 7.13: Dwell time vs. global salience scores

To gain further insight regarding this aspect, we computed the salience maps using Deep Gaze II (Kümmerer et al., 2016), a computational salience model that is not limited to low-level features, but also makes use of high-level cues, obtained by pre-training the model with image object recognition tasks. We repeated the same analyses as with the GBVS model and we found that metrics derived from Deep Gaze salience maps did have a non-zero, yet moderate correlation with our proposed global salience. This, together with previous evidence about the importance of low- and high-level features in detecting fixations (Kümmerer et al., 2017), matches our finding that global salience cannot be explained by low-level properties of the images. However, the relatively low correlation further suggests that the initial preference for one of the images does not depend only on properties of the individual salience maps.

According to previous research, initial eye movements in young adults are based on bottom-up image features, whereas socially relevant content is fixated later in time (Açık et al., 2010). Interestingly, as described above, we found that this was not the case when two images have been shown at the same time. Considering the very short reaction time between stimulus onset and the observers reaction to fixate one of the two images, it seems surprising that participants had to pre-scan both images in their peripheral visual field before initialising the first saccade. Thus, in contrast to classical salience maps, we might argue that the global salience of an image highly relates to the semantic and socially relevant content.

In order to further investigate the effects of the global image salience, we also evaluated the total time of image exploration, that is the dwell time. We hereby found that, different to the initial saccade, the spatial location of images did not affect the time participants explored the individual images of each image pair. However, the task and familiarity had an effect. We saw that in the task where participants had to select the new image, new images were explored longer than previously seen images. In contrast, the task asking to select the old image led to an almost equal exploration time on new and familiar images. Therefore, we conclude that participants in general tended to explore new images for a slightly longer time. Nevertheless and most importantly, we saw generally—and independent of the spatial location, task and familiarity—that images with higher global salience were explored longer in time. Thus, images with larger global salience did not only attract initial eye movements after stimulus onset, but also led to longer exploration times. These results support our assumption, that the global salience score of an image can also be interpreted as a measure of the general attraction of an image, in comparison to other images.

In this regard, note that although we considered the location of the first fixation as the target variable to model the global salience scores and carry out the subsequent analyses, the same computational model and procedures can be used to model alternative aspects of the behavioural responses. For instance, the model could be trained to fit the dwell time—which we have found to be positively correlated with the global salience based on the first fixations—, the engagement—

time until fixating away—or the number of saccades.

In spite of the high performance of our computational model and its potential to assign reliable global salience scores to natural images, an important limitation is that the model and thus the scores are dependent on the image set that we used. Whereas local salience maps rely on image features, our proposed global salience model relies on the differences between the stimuli and the behavioural differences that they elicit on the participants. We observed significant differences between image categories, for example humans versus indoor scenes, but this is only one initial step and future work should investigate what other factors influence the global image salience. For example, it would be interesting to train a deep neural network with a possibly larger set of images and the global salience scores learned by our model as labels, similarly to how Deep Gaze was trained to predict fixation locations. This could shed more light on what features make an image more globally salient.

Another related, interesting avenue for future work is investigating the global salience in homogeneous data sets, that is with images of similar content. Our work has shown that large differences exist between images with somehow different content, for instance containing humans or not. However, we did not observe significantly difference global salience between natural and urban scenes (see Figure 7.2b), although significant difference do exist between specific images. An interesting question is: *what* makes one image more likely to attract the first fixation, when presented alongside a semantically similar image? We think an answer to this question can be sought by combining a similar experimental setup as the one presented in this work, with additional data, and making use of advanced feature analysis, such as deep artificial neural networks, as mentioned above.

For instance, small changes in the context information of single images, might already have a dramatic influence on reaction times in decision tasks (Kietzmann & König, 2015). In addition, the global salience was based on eye movement behaviour of human data. Depending on the choice of participants, e.g. different culture, age, personal interests and emotions, our model could have revealed different results (Balcetis & Dunning, 2006; Dowiasch et al., 2015). Again, further studies might use the model on a wider range of participants, in order to validate the specific global salience and thus attraction of images.

In contrast, differences in the global salience between participant groups could be a great advantage in certain research fields. In medical applications for instance, researchers could identify specific diseases, such autistic spectrum disorder (ASD). In such example, our method could generate a model of the global visual salience of both control people and individuals with certain condition, and then be used for diagnosis. Another use case of our model would be marketing research, where the attraction of different images could be compared adequately based on intuitive visual behaviour. Thus, depending on the research question, the global image salience might provide a new insight in prediction and analysis of visual behaviour.

7.7 Conclusion

Previous research has investigated the local salience properties of single images, which has helped understand visual behaviour. However, assigning a single and unique global salience score to an image as a whole has been neglected. Here, we thus trained a logistic regression model to learn unique, global salience scores for each tested image. We hereby showed that images can indeed be ranked according to their global salience, providing a new method to predict eye movement behaviour across images with distinct semantic content. These results could be used in a variety of research, such as medicine or marketing.

Bibliography

Açık, A., Sarwary, A., Schultze-Kraft, R., Onat, S., and König, P. Developmental changes in natural viewing behavior: Bottom-up and top-down differences between children, young adults and older adults. *Frontiers in Psychology*, 2010.

Baddeley, R. J. and Tatler, B. W. High frequency edges (but not contrast) predict where we fixate: A Bayesian system identification analysis. *Vision Research*, sep 2006.

Balcetis, E. and Dunning, D. See what you want to see: Motivational influences on visual perception. *Journal of Personality and Social Psychology*, 2006.

Barton, J. J. S., Radcliffe, N., Cherkasova, M. V., Edelman, J., and Intriligator, J. M. Information processing during face recognition: The effects of familiarity, inversion, and morphing on scanning fixations. *Perception*, aug 2006.

Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 1952.

Calen Walshe, R. and Nuthmann, A. Asymmetrical control of fixation durations in scene viewing. *Vision Research*, jul 2014.

Connor, C. E., Egeth, H. E., and Yantis, S. Visual attention: bottom-up versus top-down. *Current Biology*, 2004.

Corbetta, M. and Shulman, G. L. Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, mar 2002.

Desimone, R. and Duncan, J. Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 1995.

Dowiasch, S., Marx, S., Einhäuser, W., and Bremmer, F. Effects of aging on eye movements in the real world. *Frontiers in Human Neuroscience*, feb 2015.

Egeth, H. E. and Yantis, S. Visual attention: Control, representation, and time course. *Annual Review of Psychology*, feb 1997.

Einhäuser, W., Spain, M., and Perona, P. Objects predict fixations better than early saliency. *Journal of Vision*, nov 2008.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research (JMLR)*, 2008.

Geisler, W. S. and Cormack, L. K. Models of overt attention. In *The Oxford Handbook of Eye Movements*, pp. 439–454. Oxford University Press, 2011.

Guo, K. Initial fixation placement in face images is driven by top–down guidance. *Experimental Brain Research*, jul 2007.

Guo, K., Meints, K., Hall, C., Hall, S., and Mills, D. Left gaze bias in humans, rhesus monkeys and domestic dogs. *Animal Cognition*, may 2009.

Harel, J., Koch, C., and Perona, P. Graph-based visual saliency. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2007.

Itti, L. and Koch, C. Computational modelling of visual attention. *Nature Reviews Neuroscience*, mar 2001.

Itti, L., Koch, C., and Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 1998.

Kaspar, K. What guides visual overt attention under natural conditions? Past and future research. *ISRN Neuroscience*, 2013.

Kaspar, K. and König, P. Emotions and personality traits as high-level factors in visual attention: a review. *Frontiers in Human Neuroscience*, 2012.

Kastner, S. and Ungerleider, L. G. Mechanisms of visual attention in the human cortex. *Annual Review of Neuroscience*, mar 2000.

Kietzmann, T. C. and König, P. Effects of contextual information and stimulus ambiguity on overt visual sampling behavior. *Vision Research*, may 2015.

Koch, C. and Ullman, S. Shifts in selective visual attention: Towards the underlying neural circuitry. In Vaina, L. (ed.), *Matters of Intelligence*, chapter 4, pp. 115–141. Springer, Dordrecht, 1987. doi: 10.1007/978-94-009-3833-5_5.

Kollmorgen, S., Nortmann, N., Schröder, S., and König, P. Influence of low-level stimulus features, task dependent factors, and spatial biases on overt visual attention. *PLOS Computational Biology*, may 2010.

König, P., Wilming, N., Kietzmann, T. C., Ossandón, J. P., Onat, S., Ehinger, B. V., Ramos Gameiro, R., and Kaspar, K. Eye movements as a window to cognitive processes. *Journal of Eye Movement Research*, 2016.

Kowler, E. Eye movements: The past 25 years. *Vision Research*, 2011.

Kümmerer, M., Wallis, T. S., and Bethge, M. DeepGaze II: Reading fixations from deep features trained on object recognition. *arXiv preprint arXiv:1610.01563*, 2016.

Kümmerer, M., Wallis, T. S., Gatys, L. A., and Bethge, M. Understanding low-and high-level contributions to fixation prediction. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

Liversedge, S. P. and Findlay, J. M. Saccadic eye movements and cognition. *Trends in Cognitive Sciences*, jan 2000.

Luce, R. D. *Individual Choice Behavior: A Theoretical Analysis*. Dover Publications, Inc., Mineola, New York, dover edit edition, 2005.

Niebur, E. and Koch, C. Control of selective visual attention: Modeling the "where" pathway. *Advances in Neural Information Processing Systems (NeurIPS)*, 1996.

Ossandón, J. P., Onat, S., and König, P. Spatial biases in viewing behavior. *Journal of Vision*, 2014.

Ramos Gameiro, R., Kaspar, K., König, S. U., Nordholt, S., and König, P. Exploration and Exploitation in Natural Viewing Behavior. *Scientific Reports*, dec 2017.

Rauthmann, J. F., Seubert, C. T., Sachse, P., and Furtner, M. R. Eyes as windows to the soul: Gazing behavior is related to personality. *Journal of Research in Personality*, apr 2012.

Reinagel, P. and Zador, A. M. Natural scene statistics at the center of gaze. *Computation in Neural Systems*, 1999.

Riche, N., Duvinage, M., Mancas, M., Gosselin, B., and Dutoit, T. Saliency and human fixations: State-of-the-art and study of comparison metrics. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

Spalek, T. M. and Hammad, S. The left-to-right bias in inhibition of return is due to the direction of reading. *Psychological Science*, jan 2005.

Tatler, B. W. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, nov 2007.

Tjur, T. Coefficients of determination in logistic regression models—a new proposal: The coefficient of discrimination. *The American Statistician*, 2009.

Wadlinger, H. A. and Isaacowitz, D. M. Positive mood broadens visual attention to positive stimuli. *Motivation and Emotion*, mar 2006.

Zaeinab, A., Ossandón, J. P., and König, P. The dynamic effect of reading direction habit on spatial asymmetry of image perception. *Journal of Vision*, 2016.

Chapter 8

Image identification from brain activity

Contributors

Wietske Zuiderbaan, Ben M. Harvey and Serge O. Dumoulin are the authors of the article upon which this work builds. Wietske and Serge supervised the work during my internship at their lab. Akhil Edadan and Peter König participated in the discussions.

Outreach

This chapter extends the following publications:

- *Saliency and the population receptive field model to identify images from brain activity.* **Alex Hernández-García**, Wietske Zuiderbaan, Akhil Edadan, Serge O. Dumoulin, Peter König. Annual Meeting of the Visual Sciences Society (VSS, poster presentation), 2019.

The goal of computational neuroscience is to understand the mechanisms and principles by which the brain yields behaviour and cognition. One of the approaches to deepen our understanding of the brain is to develop models that can make predictions about brain activity. These methods, sometimes referred to as *brain/mind reading*, not only have some practical applications, but they also provide insights about the underlying neural processes that enable the prediction (Tong & Pratte, 2012).

One particular example for the study of the visual system is the identification of presented images from brain imaging recordings, such as fMRI, in the visual cortex. While a successful approach to this challenge is to use statistical models that can learn activations patterns from a set of recordings (Kay et al., 2008), an alternative approach is to rely on biologically inspired encoding models. A recent proposal of this type showed that it is possible to identify the presented stimulus from a set of natural images by comparing the measured activity on areas V1, V2 and V3 with a predicted response profile encoded by the low-parametric population receptive field (pRF) model and contrast information from the images (Zuiderbaan et al., 2017). In the work presented in this chapter, we follow up this methodology by studying the predictive power of salience information from the images—instead of contrast—combined with the pRF model, and extend the analysis to higher visual areas: V1, V2, V3, hV4, LO12 and V3AB[1].

[1]This work was carried out during a 2-months internship at the Spinoza Centre for Neuroimaging in Amsterdam, in early 2018, supervised by Dr. Wietske Zuiderbaan and Professor Serge Dumoulin. Besides the potential scientific interest of the work, which was presented as a poster at the 19th Annual Meeting of the Vision Sciences Society (VSS), the internship was part of the training programme of my PhD fellowship, a Marie Skłodowska-Curie Innovative Training Network. A requirement of the programme is to carry out interdisciplinary internships at laboratories that are part of the partnership. As an additional contribution of my internship, I developed interactive visualisation tools in Python to better interpret the kind of data used in this project which became part of the laboratory's repository.

8.1 Methods

The goal of this study was to assess whether the salience information of images is predictive of brain activations in the visual cortex elicited by the visualisation of natural images. As a follow-up study of the work by Zuiderbaan et al. (2017), a significant part of the methodology is borrowed from the first publication and, in particular, the neuroimaging data is the same. In this section, we will outline the most relevant aspects of the methodology that is not original of this work and refer to the original publication for further details, so as to focus on the aspects that are novel here.

8.1.1 Visual stimuli

As a data set of images we used a subset of 45 images from the Berkeley Segmentation Dataset and Benchmark (Martin et al., 2001), which was already used in some of the experiments by (Zuiderbaan et al., 2017). The data set was processed such that the resulting images had a resolution of 538×538 pixels and were applied a circular fading mask. The set of 45 images as they were used in the experiments are shown in Figure 8.1.

Figure 8.1: The set of 45 natural images used in the experiments

8.1.2 Functional imaging

The brain activations in the visual cortex were acquired in a neuroimaging study with functional magnetic resonance imaging (fMRI), carried out originally for the work presented by Zuiderbaan et al. (2017). Therefore, the details can be found in that first article and here we will summarise the most relevant information: Two participants (one female; ages 28-38 years) took part in the fMRI study, were their brain responses to the natural images were collected. Each image was shown 18 times for a duration of 300 ms, with a radius of 5.5° from the participant point of view. The neural response at each cortical location (voxel) was obtained through standard GLM-analysis (Friston et al., 1995): each voxel was summarised by the t-value of the fit between the predicted time series—the stimulus presentation convolved with the hemodynamic response function—and the measured BOLD signal.

In addition to the voxel responses to each stimulus, the optimal parameters of the population receptive field (pRF) model were estimated for each participant using the standard approach described by Dumoulin & Wandell (2008): For each voxel, the model estimates the position of the receptive field (x, y) and the size σ (standard deviation) using a Gaussian kernel. The parameters of the pRF were estimated using conventional contrast-defined moving-bar apertures with natural image content from images of the same data set but distinct from the set of 45 images used in the image identification experiments.

Zuiderbaan et al. (2017) analysed areas V1, V2 and V3 of the visual cortex as regions of interest. Here, we extended the analysis to higher visual areas: on the lateral occipital complex (LO-1 and LO-2: LO12), on the ventral occipital (hV4) and on the dorsal occipital (V3A and V3B: V3AB) (Wandell et al., 2007). See Figure 8.2 for an illustration of the regions of interest in the human visual cortex.

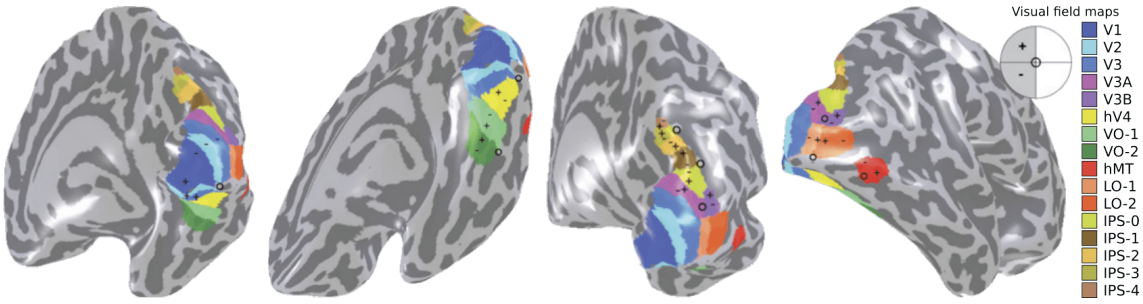


Figure 8.2: Adapted from (Wandell et al., 2007) with permission from the authors: Visual field maps in the human visual cortex. In our analyses, V3A and V3B were considered a single region of interest (V3AB), and so were LO-1 and LO-2 (LO12).

8.1.3 Salience and contrast maps

One of the goals of this work was to contrast the correlation of salience maps with the activations in the visual cortex, and the predictive ability of contrast maps—calculated as the root mean squared contrast—which was assessed in the original article by Zuiderbaan et al. (2017). The choice of contrast is motivated by the evidence that the early visual cortex responds strongly to differences in contrast (Boynton et al., 1999; Olman et al., 2004). However, the visual cortex certainly responds to other properties of the stimuli and hence the interest in studying salience.

In order to analyse the correlation of salience with the activations in the visual cortex, we computed the salience maps of each image using two distinct salience models. One of the models is DeepGaze II[2], the then state-of-the-art salience model for various metrics. The second model is ICF, which stands for Intensity Contrast Features, proposed by Kümmerer et al. (2017). ICF is trained on the same readout neural network as DeepGaze, but instead of using the VGG features, the input to the network are 5 intensity and 5 contrast feature maps.

Our choice of salience models, out of the many models proposed in the vast literature on image salience, was motivated by various reasons. First of all, we chose DeepGaze as it was the state-of-the-art model in various metrics of image salience evaluation (Kümmerer et al., 2016) and considering a model that accurately predicts the salient regions of images is also desirable for studying how salience information is encoded in the visual cortex. Nonetheless, as we have discussed in Chapter 7, visual attention is a complex brain mechanism and salience maps capture different aspects of it. For instance, we have discussed how visual attention can be guided by both bottom-up factors, as well as top-down factors and higher-level features. This motivated the

[2]Note that in Chapter 7 we also used DeepGaze to analyse whether our proposed global salience was related or independent from the local salience properties of images, proposed by Kümmerer et al. (2016). DeepGaze II—for better readability, in what follows we will simply refer to it as DeepGaze—is a model that uses the features extracted by a deep neural network, VGG (Simonyan & Zisserman, 2014), trained on image object categorisation tasks as inputs to a 4-layer readout neural network optimised for salience prediction

inclusion of a model that is limited to lower-level—intensity and contrast—features, in this case ICF. The fact that DeepGaze and ICF share part of the architecture facilitates the comparisons.

Interestingly, Kümmerer et al. (2017) analysed and compared the properties of DeepGaze and ICF, and found that while DeepGaze generally outperformed ICF in terms of the evaluation metrics used to assess salience models, ICF was more accurate in a large number of images. In particular, DeepGaze succeeds at predicting salient regions associated with higher-level factors such as objects and faces, while ICF was superior in the cases were fixations are driven by, for instance, high contrast. By way of illustration, in Figure 8.3 we show the DeepGaze and ICF salience maps of three images from the data set, as well as their contrast maps, which were used in the original study.

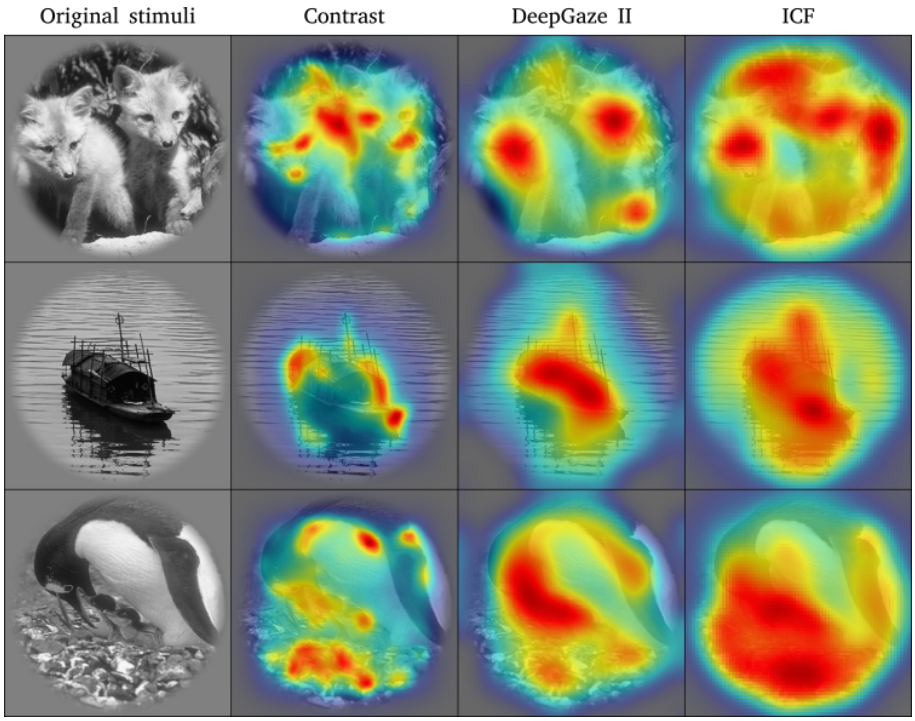


Figure 8.3: Contrast, DeepGaze II and ICF maps of three images of the data set.

8.1.4 Brain response predictions

To assess the correspondence between the salience and contrast maps and the brain activations, we first calculated a prediction response profile of every image as the summed overlap of the map with the pRF weighting function of each cortical location, normalised by the total volume of the pRF:

$$p = \frac{\sum_{i=1}^{N} w_i S_i}{\sum_{j=1}^{M} w_j} \tag{8.1}$$

where S_i denotes the value of the feature map S—either DeepGaze, ICF or contrast—normalised as a probability distribution, at pixel i; N is the number of pixels within the window of the pRF and M is the total number of pixels in the stimulus area. w_i is the pRF weighting function, whose parameters were obtained as described in Section 8.1.2 and in (Zuiderbaan et al., 2017) in more detail:

$$w_i = \exp\left(-\frac{(x_i - x_c)^2 + (y_i - y_c)^2}{2\sigma^2}\right) \tag{8.2}$$

where x_i and y_i are the locations of pixel i; x_c and y_c are the centre of the pRF in the visual field; and σ is the size of the Gaussian kernel of the pRF. Note that in (Zuiderbaan et al., 2017) this is the procedure used to compute the predictions of synthetic images, while for natural images a different method was used, which was specific for the way the contrast information was obtained. Here, we use this method, since we can generalise the prediction p in Equation 8.1 to any feature map S—DeepGaze, ICF or contrast.

8.1.5 Evaluation metrics

Let us formally express the measurements and predictions taking all variables into account—feature maps, areas, images, voxels—in order to describe the evaluation metrics we used in our analysis. First note that, as in the original study by Zuiderbaan et al. (2017), we only considered for the analysis the voxels with positive t-value, pRF eccentricity values within 0.5—4.5° and pRF variance explained larger than 55 %. For every image k and every visual area A studied—V1, V2, V3, hV4, LO12 and V3AB—we have a *measured* response profile

$$\mathbf{m}_k^A = m_{k,1}^A, \ldots, m_{k,D_A}^A$$

where D_A is the number of (*valid*) voxels in area A. Correspondingly, we have the *predicted* response profiles for every image k, every visual area A and by every feature map S, that is DeepGaze, ICF and contrast:

$$\mathbf{p}_k^{A,S} = p_{k,1}^{A,S}, \ldots, p_{k,D_A}^{A,S}$$

where every $p_{k,v}^{A,S}$ is computed as in Equation 8.1. As a similarity metric we compute the Pearson correlation between measured responses and predicted profiles. We will denote by $r_{k,l}^{A,S}$, or simply $r_{k,l}$ abusing notation, to the correlation between the measured response of image k, $\mathbf{m}_k^A$, and the predicted profile of image l, $\mathbf{p}_l^{A,S}$:

$$r_{k,l}^{A,S} = corr(\mathbf{m}_k^A, \mathbf{p}_l^{A,S}) \tag{8.3}$$

In order to assess the image identification accuracy we simply considered a correct identification if $r_{k,k} > r_{k,l}, \forall\, k \neq l$. Additionally, as a more informative metric, we combined the correlation values into a confidence score that represents how hard it is to distinguish the actual presented image k from the other candidate images in the data set, based on the correlation values:

$$c_k^{A,S} = r_{k,k}^{A,S} - \frac{1}{K}\sum_{l=1}^{K} r_{k,l}^{A,S} \tag{8.4}$$

Finally, in order to have a compact measure to compare the predictivity of the different feature maps on every visual area, we also performed a representational similarity analysis[3] (RSA) (Kriegeskorte et al., 2008). In order to perform RSA, we constructed representational dissimilarity matrices (RDM) for both the measured responses and the predicted profiles, $M_{k,l}^A$ and $P_{k,l}^{A,S}$ respectively, where each entry (k, l) of the matrices is 1 minus the Pearson correlation between the profile for image k and image j. In this case, the correlation is not computed between the measured and the predicted profiles, but between two measured responses for $M_{k,l}$ and two predicted profiles for $P_{k,l}$. Thus, the RDMs are symmetric and the values in the diagonals are zero. As a summary metric to compare M and P we computed the Kendall correlation.

[3] In Chapter 5, we also used representational similarity analysis to compare the features learnt by artificial neural networks and the representations measured in the inferior temporal cortex.

8.2 Results and discussion

We first analyse the distribution of the confidence of the predictions for each model—feature map—and visual area, shown in Figure 8.4. Although the results are complex and not highly consistent, several interesting conclusions can be drawn. First, we observe that the identification ability of all models is best on V1 and decreases in higher visual areas. This was observed by (Zuiderbaan et al., 2017) too, who analysed V1, V2 and V3; and we here confirmed it, although we had hypothesised that salience maps might be more discriminative in higher visual areas. Unfortunately, it is not possible to conclude from our results that contrast and salience are less predictive of the activations in higher visual areas because this result could be also explained by the increased pRF sizes (Smith et al., 2001). In what follows, our analysis will focus in the earlier areas—V1, V2, V3 and, to a lesser extent, hV4—where the identification performance is better and the differences between models larger.

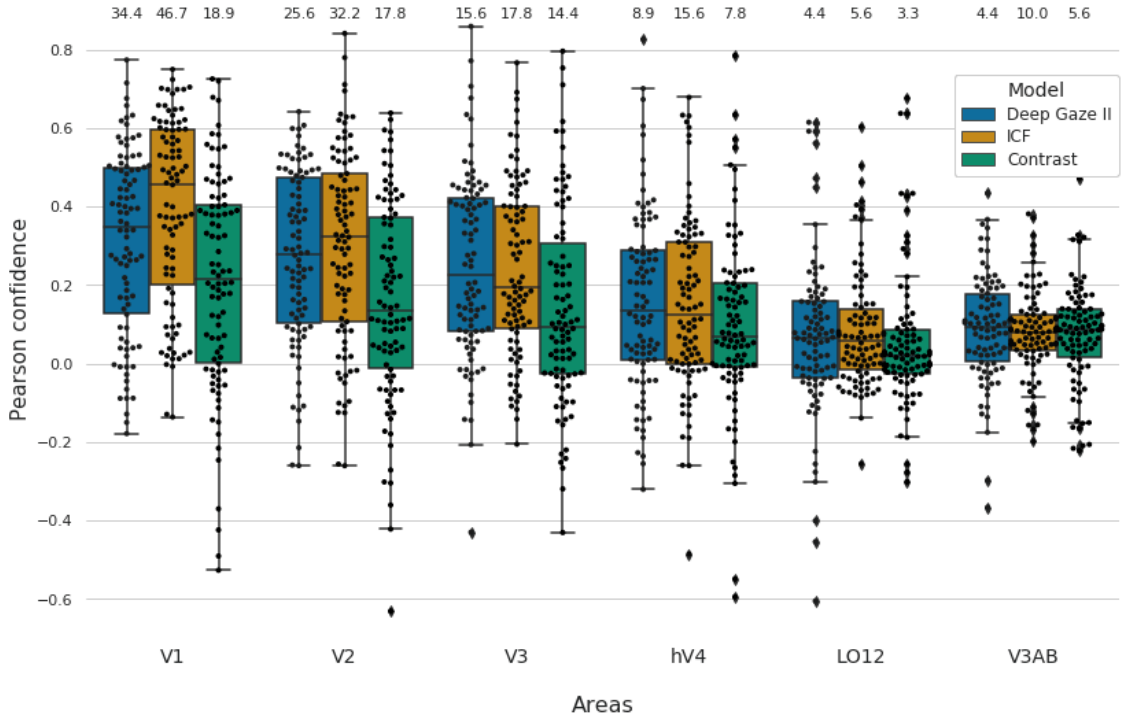


Figure 8.4: Distribution of the confidence values (see Equation 8.4) for every feature map and visual area. Each black dot corresponds to the confidence of the prediction for one image k, from the data of one of the experimental participants. The numbers over each box represent the identification accuracy (as a percentage).

Second, the most relevant observation is that both salience models, DeepGaze and ICF, seem more predictive of the measured activations in the visual cortex than the contrast maps of the images. A significant difference can be observed in both the median of the distribution of the confidence values and in the identification accuracy, shown over each box. For instance, the identification accuracy on V1 using DeepGaze and ICF maps is 34.4 % and 46.7 % respectively, while it is below 20 % using contrast[4].

Although image identification from brain activity has been shown before (Kay et al., 2008), the identification performance of the methods presented here is remarkable due to the simplicity of

[4]In the first publication by Zuiderbaan et al. (2017) the prediction model for natural images using contrast is computed differently and the evaluation procedures are not the same, hence the results differ slightly to the ones presented here. Nonetheless, in our study we analysed both methods and the results are qualitatively very similar. We here present the results using contrast maps, since the method is identical for salience maps.

the model. The main differences between this pRF-based method and the model presented by Kay et al. (2008) were reviewed by Zuiderbaan et al. (2017), but most importantly, note that here we estimated only 3 pRF parameters for each voxel—location and size of the receptive field—which were obtained from the fMRI recordings of a standard pRF session: responses to conventional moving-bar apertures during about half an hour. Then, the 3 pRF parameters are combined with the contrast or salience maps of the images to predict the activity of each voxel. In contrast, Kay et al. (2008) recorded the activations to 1,750 natural images (about 5 hours of scanning time) to fit a Gabor-Wavelet-Pyramid model with 2,730 parameters per voxel. Summarised, image identification with the pRF model requires a short scanning time and does not require training a model with brain activations towards natural image. Thus, it can be easily applied to any other image.

The fact that identifying natural images from brain activity using salience maps is more accurate than using contrast information from the images tells us that the image salience *under* the receptive field of each voxel is more discriminative than the contrast. This may come as a surprise, since the early visual cortex is well known to respond strongly to differences in contrast (Boynton et al., 1999; Olman et al., 2004). Salience maps also certainly contain contrast information, in our analysis especially the maps obtained with the ICF model, but not only. Therefore, our results can be interpreted as additional evidence of the activations in the early visual cortex being shaped by multiple factors, likely top-down influences (Treue, 2003).

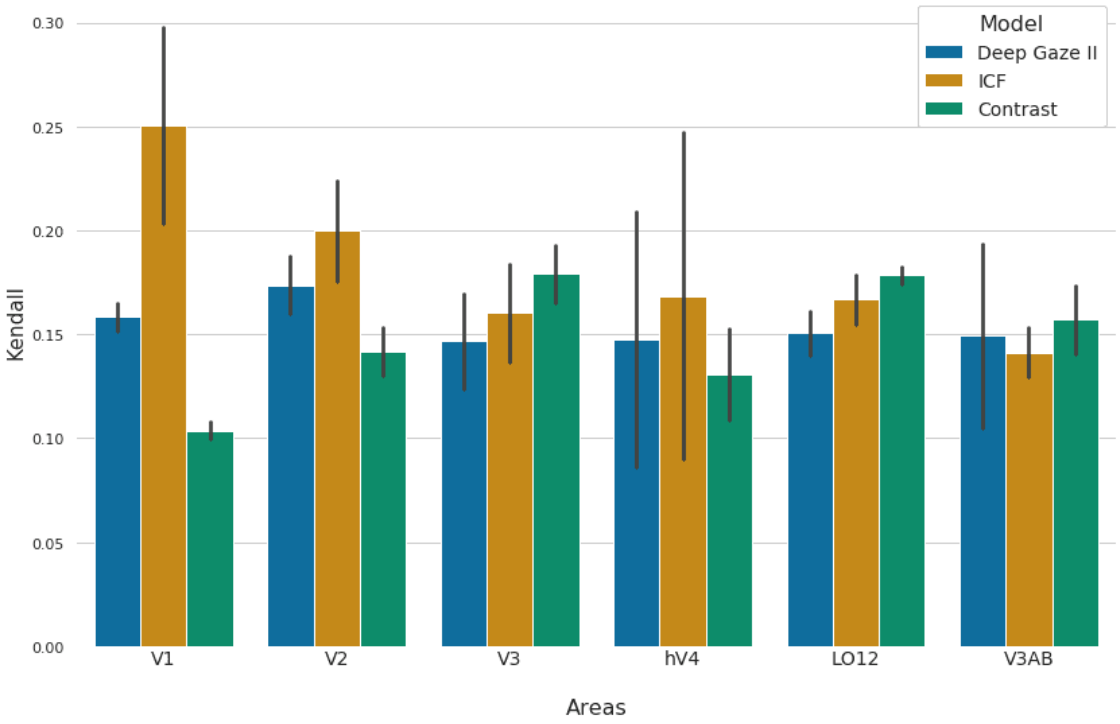


Figure 8.5: Results of the representational similarity analysis (RSA): Kendall correlation between the representational dissimilarity matrices of the measured responses and the predicted profiles. The error bars correspond to the variation across experimental participants.

Another interesting conclusion derives from the comparison between the two salience models analysed. Recall that DeepGaze computes the salience map by first extracting features through a deep neural network trained on image object recognition tasks. Therefore, these features likely contain high-level information such as face and object detectors and the model is particularly accurate at spotting salience driven by such factors, as reflected in Figure 8.3. On the contrary, ICF does not extract high-level features but is restricted to intensity and contrast features. Here, we found that image identification is more accurate by using ICF than DeepGaze salience maps. While this is observable in Figure 8.4, we additionally perform a representational similarity ana-

lysis (RSA) to derive a compact metric of the ability of each model to discriminate the brain activations of each image. We present these results in Figure 8.5, which confirms the conclusion that ICF is more discriminative than DeepGaze in the earlier visual areas.

One more way to visualise the superiority of ICF at predicting brain activations in the early visual cortex is by directly analysing the correlation matrices. In Figure 8.6, we plot the correlation matrices of each model on V1, where each element in the matrix encodes the correlation between the measured responses of one image (rows) and the predicted profile of another image (columns), as in Equation 8.3. Note that a correct identification occurs when the element in the diagonal—correlation between measured and predicted profile of the same image—has the highest value in the row. Visually, it becomes apparent that the diagonal in the ICF model has higher values and is better discriminated from the rest of the matrix, than in the DeepGaze model, and yet more in the contrast model. In view of these results, we conclude that image salience is more predictive of the brain activity in the early visual cortex than contrast information, and hypothesise that ICF may be more accurate than DeepGaze because the salience of the latter is driven by high-level information that may not correlate with the activity in the early areas of the visual cortex.

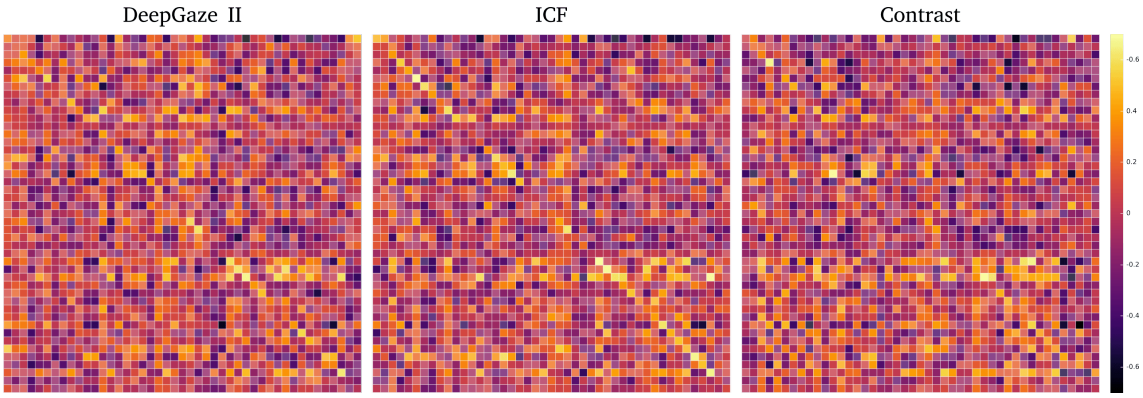


Figure 8.6: Correlation matrices of the three models for the predictions in visual area V1. Each entry (i, j) of the matrices represent $r_{i,j}^{V1,S} = corr(\mathbf{m}_i^{V1}, \mathbf{p}_j^{V1,S})$.

The data we obtained from the predictions allows for multiple levels of analysis, since there are several factors at play: three feature maps, six visual areas, two subjects, etc. During the two-months internship in which this project was carried out, I developed multiple interactive visualisation using Bokeh[5] and the interactive functionality of Jupyter notebooks. Interactive visualisation provides insights that are hardly accessed otherwise and is useful to guide the more systematic, conventional, statistical analyses. In particular, it can be used to analysed the results at different levels of analysis and to look into specific details or data points. Some examples are shown in Figure 8.7.

Besides the main conclusions discussed above, some other observations are the following: In general, we found a positive linear correlation in the prediction confidence between the predictions on the data from both participants in areas V1, V2, V3 and hV4 ($r = 0.76, 0.80, 0.78, 0.71$ respectively) and less so in LO12 and V3AB ($r = 0.42, 0.17$), where the predictions were also less confident. We also found a correlation between the predictions across visual areas, that is images that were confidently predicted in V1 were also in V2 ($r = 73$) and in V3 ($r = 0.56$), and the correlation decreases further in higher visual areas, as expected.

The prediction confidence was also positively correlated between DeepGaze and ICF ($r = 0.58, 0.62, 0.50, 0.58$ in areas V1, V2, V3 and hV4), although with exceptions, that is several images were confidently predicted by ICF but not by DeepGaze, and vice versa. The interactive visualisation was useful to identify and study these cases. Not surprisingly, the correlation was much lower between the salience and the contrast models. For instance, in V1, the correlation between DeepGaze and contrast and ICF and contrast was $r = 0.25, 0.35$, respectively.

[5]Bokeh is an interactive visualisation library for Python: www.bokeh.org

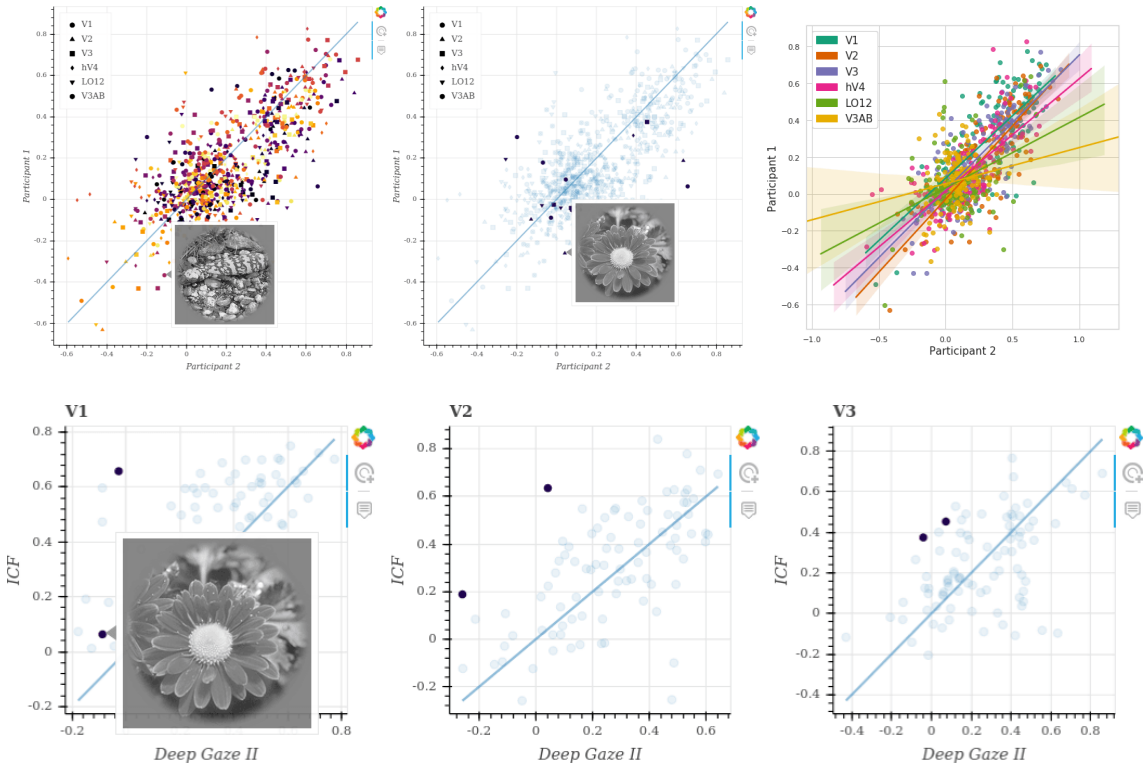


Figure 8.7: Examples of interactive scatter plots of the Pearson confidence (Equation 8.4) to contrast different experimental variables. The interactive plot allows to hover over the data points to visualise the image that they represent and highlight all data points of one image by clicking on them. Bottom row: DeepGaze vs. ICF on areas V1, V2 and V3. Top row: comparison of the confidence values of the predictions from the data of the two experimental participants. From left to right: data from the three models, represented with different shapes for each visual area and different colours for each image; highlight of one image, after clicking on one point; linear fit for each visual area separately.

8.3 Conclusion

In this chapter we have presented the results of a project carried out during a two-months internship at the Spinoza Centre for Neuroimaging in Amsterdam, in which we extended the work presented by Zuiderbaan et al. (2017). In particular we analysed the discriminability of salience maps to identify images from brain activity in the visual cortex, using a low-parametric model of the receptive field of the measured cortical locations, the population receptive field (pRF) model (Dumoulin & Wandell, 2008).

We contrasted the identification performance of two distinct salience models—DeepGaze and ICF— and the contrast-based model originally presented by Zuiderbaan et al. (2017) and found that the salience information within the receptive fields of the voxels is more predictive of the brain activity than the contrast information. Furthermore, we observed that ICF, whose salience maps are computed using low-level features, is more predictive than DeepGaze, which encodes high-level information such as salience driven by faces and objects.

Overall, this analysis demonstrates that this simple method of prediction of brain activity using the pRF model can be extended to other types of image information beyond contrast, enabling further analysis. Future work may use this kind of analysis to further understand the computations in the early visual cortex and extend this methodology to other cortical areas.

Bibliography

Boynton, G. M., Demb, J. B., Glover, G. H., and Heeger, D. J. Neuronal basis of contrast discrimination. *Vision Research*, 1999.

Dumoulin, S. O. and Wandell, B. A. Population receptive field estimates in human visual cortex. *Neuroimage*, 2008.

Friston, K. J., Holmes, A. P., Poline, J., Grasby, P., Williams, S., Frackowiak, R. S., Turner, R., et al. Analysis of fmri time-series revisited. *Neuroimage*, 1995.

Kay, K. N., Naselaris, T., Prenger, R. J., and Gallant, J. L. Identifying natural images from human brain activity. *Nature*, 2008.

Kriegeskorte, N., Mur, M., and Bandettini, P. A. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2008.

Kümmerer, M., Wallis, T. S., and Bethge, M. DeepGaze II: Reading fixations from deep features trained on object recognition. *arXiv preprint arXiv:1610.01563*, 2016.

Kümmerer, M., Wallis, T. S., Gatys, L. A., and Bethge, M. Understanding low-and high-level contributions to fixation prediction. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

Martin, D., Fowlkes, C., Tal, D., and Malik, J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *IEEE International Conference on Computer Vision (ICCV)*. 2001.

Olman, C. A., Ugurbil, K., Schrater, P., and Kersten, D. Bold fmri and psychophysical measurements of contrast response to broadband images. *Vision Research*, 2004.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Smith, A. T., Singh, K. D., Williams, A., and Greenlee, M. W. Estimating receptive field size from fmri data in human striate and extrastriate visual cortex. *Cerebral Cortex*, 2001.

Tong, F. and Pratte, M. S. Decoding patterns of human brain activity. *Annual Review of Psychology*, 2012.

Treue, S. Visual attention: the where, what, how and why of saliency. *Current Opinion in Neurobiology*, 2003.

Wandell, B. A., Dumoulin, S. O., and Brewer, A. A. Visual field maps in human cortex. *Neuron*, 2007.

Zuiderbaan, W., Harvey, B. M., and Dumoulin, S. O. Image identification from brain activity using the population receptive field model. *PLOS ONE*, 2017.

Chapter 9

General discussion

In this dissertation I have presented a series of experimental studies and discussions revolving around machine learning for image understanding, visual perception and visual neuroscience. An overarching objective of this work was to explore and exploit the connections between these fields, combining the tools and techniques common to each discipline.

A central subject of the dissertation has been data augmentation. Data augmentation has been ubiquitously used to train machine learning models on image tasks since the early 1990s, but it has received little scientific attention. In the first part of the thesis, we tried to bring data augmentation to the fore and study its role as implicit regularisation of machine learning algorithms and its potential to incorporate inductive biases from visual perception and biological vision. While on the surface data augmentation is just a method to synthetically increase the number of examples in a data set, we have here analysed it as a technique that encodes effective priors from perception: The image transformations typically included in data augmentation techniques—rotations, translations, scaling, changes in illumination, etc.—coincide with those that are plausible in the real world as we perceive it. Likely not by coincidence, the visual cortex of our brains represents objects under these transformations in a largely robust way.

From a machine learning point of view, data augmentation can be seen as a form of regularisation, in that it helps improve generalisation. Nonetheless, we discussed an important distinction between the type of regularisation provided by data augmentation—implicit regularisation—and explicit regularisation techniques (Chapter 3). The terms explicit and implicit regularisation have appeared frequently in the deep learning literature, but no formal definition had been provided, to the best of our knowledge. Hence, the terms have been in used in an inconsistent and subjective manner. We here provided formal definitions of the two concepts based on their effect on the representational capacity of the model they are applied on, alongside several examples of each category for illustration. Importantly, we argued that data augmentation does not reduce the representational capacity and therefore is not explicit but implicit regularisation. We hope our definitions find consensus in the machine learning community and foster more rigorous discussions about regularisation.

In Chapter 4, we delved into the distinction between data augmentation and explicit regularisation. We departed from the hypothesis that data augmentation improves generalisation by increasing the number of training examples through transformations that resemble those that can be find in the real world, while explicit regularisation *simply* relies on the inductive bias that simpler models should generalise better. Although this inductive bias is at the root of the feasibility of learning from data and has proven effective in uncountable applications, the prior knowledge encoded by data augmentation seems intuitively more effective. Accordingly, we challenged the need for explicit regularisation techniques such as weight decay and dropout to train deep neural networks, provided data augmentation is also employed. If large networks with orders of magnitude more learnable parameters than training examples are able to generalise well, is it neces-

sary to constrain their representational capacity? We first derived some theoretical insights from the literature that suggest that weight and dropout can be seen as *naive* data augmentation, that is without domain knowledge. We then confirmed through an empirical evaluation that models trained with data augmentation alone outperform the combination of explicit regularisation and data augmentation typically used in practice.

Although the experimental setup of our empirical study included several network architectures and data sets, with results of over 300 trained models, extended experimentation would be of course desirable. All the experimental results from training neural networks presented in this thesis have been conducted with one—occasionally two—graphical processing unit (GPU) available. It would be highly beneficial if researchers without such computational limitations extended this analysis to confirm or reject our conclusions, and therefore we made the code available alongside the publications. Another desirable extension of this part of the dissertation would be to compare data augmentation and explicit regularisation in other data modalities beyond natural images, such as speech, text or medical images.

Since one of the motivations for analysing image data augmentation was its connection with visual perception and biological vision, we hypothesised that larger variation in the image transformations seen by a neural network may induce better representational similarity with the inferior temporal cortex. This is the region in the visual cortex where it is possible to decode object classes from measured activations and invariance to transformations has been repeatedly observed. In Chapter 5, we used representational similarity analysis to compare the features learnt by artificial neural networks and the activations measured in the inferior temporal cortex through fMRI. As hypothesised, we found that models trained with heavier transformations exhibit higher similarity with the visual cortex. This study was the result of a short collaboration in which we tested the idea with a limited experimental setup. Therefore, it would also be desirable to find more evidence of our conclusion in future work, as well as delving into what specific transformations drive invariance in the higher visual cortex.

The last chapter of the block on data augmentation made the connection with visual perception and biological vision more explicit. We departed from the idea that simply applying transformations to the input images and optimising a neural network for classification may not be enough to learn robust features as in the higher visual cortex. We first observed that useful information is lost in the way data augmentation is commonly applied: every time an image is transformed according to a data augmentation scheme, it is fed into the network to compute the classification loss just as any other new image. The transformed image is not just one more image, but a perceptually plausible transformation of another image in the set. With the standard classification objectives this potentially valuable information is simply lost. Could it not be used as an inductive bias?

In order to further exploit the potential inductive bias of data augmentation, in Chapter 6 we proposed *data augmentation invariance*, a simple learning objective inspired by the increasing invariance to identity-preserving transformations observed in the ventral visual stream. Data augmentation invariance combines several novel aspects: First, we perform data augmentation within the training batches, that is we construct the mini-batches by including M transformations of each image. In this way, the model has access to the multiple transformations of an example at once—instead of separated by many iterations—and it potentially reduces the variance of the gradients. Second, we proposed a contrastive loss term that encourages similar representations of images that are transformations of each other. This has been suggested to be a key property of the inferior temporal cortex. Third, we define the data augmentation invariance objective in a layer-wise fashion, that is the representational invariance is optimised at multiple layers of the network. However, we distribute the weights of the loss terms of each layer exponentially along the hierarchy. This aimed to loosely mimic the increasing invariance along the visual cortex. We trained several architectures with data augmentation invariance and the models effectively and efficiently learnt robust representations, without detriment of the classification performance. In contrast, the representations of models trained with the standard categorical cross-entropy loss did not become more invariant to transformations than at the pixel space, in spite of being exposed to data augmentation during training.

Although our results were remarkably consistent across architectures and data sets, future work should find more evidence for the benefits of data augmentation invariance. Furthermore, we are interested in exploring other potential benefits of training with this objective. In particular, we would like to test the representational similarity of the learnt features with the inferior temporal cortex, which inspired this approach. Furthermore, it would be interesting to study whether encouraging invariance to some transformations—rotations, translations, illumination changes—induces invariance to other transformations, such as occlusions as in cutout augmentation.

In the second part of the dissertation, we moved the focus from data augmentation and artificial neural networks to visual attention and salience, using tools of cognitive science and neuroscience, such as eye-tracking and neuroimaging. In Chapter 7, we proposed and analysed the concept of *global visual salience*. While a large body of scientific literature has studied visual attention and the salience properties of images, it has mostly focused in analysing what parts and features of an image drive eye movements and are more likely to attract fixations. Here, we studied the likelihood of natural images as a whole to attract the initial fixation of a human observer, when presented in competition with other images. For this purpose, we carried out an eye tracking experiment in which we showed participants pairs of images side by side. We trained a simple machine learning algorithm with the behavioural data from the experiment and found that it is possible to predict the direction of the first saccade—left or right—given a pair of images from the data set. This implies that some images have a higher *global visual salience* than others. Specifically, faces and images with social content are most likely to be fixated first. Importantly, we also found that global salience is largely independent from the local salience properties of the images.

We believe our experimental data can be further used to study aspects of human visual attention of competing stimuli, since we mostly focused on the direction of the first fixation upon stimulus presentation. Therefore, we open sourced the data and the code of our analyses. In particular, it would be interesting to study the reaction times and engagement with the stimuli during the duration of the trials so as to find if there exist differences depending on the nature of the two images, for instance. Another interesting direction would be to more deeply study the visual properties of the images and find out whether it is possible to predict the global visual attention of novel images. Further, we hypothesised that global salience could be used as a tool or metric to better understand the visual attention behaviour of humans with conditions such as the autism spectrum disorder.

Finally, in Chapter 8, we analysed the relationship between the local salience maps of natural images and the brain activations in the early visual cortex. In particular, we followed up a previous study that demonstrated the possibility to identify natural images from brain activity using the low-parametric population receptive field (pRF) model and contrast information from the images. In our work, we extended that study by analysing the discriminability of salience maps. We compared contrast and salience maps computed with two distinct image salience models, one based on low-level features and the other based on high-level features learnt by a deep neural network. We found that salience, especially based on low-level features, is significantly more predictive of brain activity than contrast. This suggests that the activations in the early visual cortex contain information about various properties of the images, likely driven by feedback connectivity from higher areas. Moreover, the results in this chapter provided additional evidence for the possibility of studying properties of the visual cortex through predictive models based on simple tools such as the pRF model.

Before concluding this dissertation, I would like to briefly discuss some ethical considerations and the societal and environmental impact of the work presented here. First, although compared to much of the deep learning literature the computational resources used for this work were small, some of the results of this thesis required training multiple neural network models, especially for Chapter 4. Training these models certainly contributed negatively on the environment with emmision of carbon dioxide, as reported in the chapter. In order to disseminate my work and engage with other scientists, I travelled by plane to attend several conferences, which also had a negative impact on climate change. I strongly advocate minimising the impact of scientific activity on the environment. One way of positively contributing to reduce this impact is through data sharing. Hence, we have made available much of the data collected for this work, which will also hope-

fully contribute to more open science. Currently, deep neural networks are remarkably energy-inneficient compared to brains. Incorporating better inductive biases, as we have discussed in this thesis, may contribute to more efficient machine learning algorithms. Second, while I do not envision a direct negative use of the work presented here, I believe that as work that aims to advance our technology, it has the potential of being misused or negatively impact our society. As Professor Ruha Benjamin puts it, "technology can exclude without being explicitly designed for it". I hope this is not the case of my work and I explicitly disapprove the use of the results, conclusions, data and code related to this work for applications that incite racism, sexism or unequal treatment of marginalised groups.

In sum, in this dissertation we have presented the results of various projects connecting different fields, such as machine learning, cognitive science and computational neuroscience. While science clearly needs the depth of very narrow studies, we have here tried to show the supplementary value of an interdisciplinary approach to science. In particular, I believe that understanding the nature of learning systems—both algorithms and brains—requires the collaboration of scientists of multiple disciplines, as many other researchers have argued before me. Learning algorithms will become more effective and efficient by incorporating insights from the brain; and we will deepen our understanding of the brain by using the tools of improved machine learning.

"Many small people
who in many small places
do many small things
can change the face the world"