University of Amsterdam

Artificial neural networks as models of information processing in biological neural networks

Literature thesis by Andreas Wolters

Student number: 11 11 97 64

res.M.Sc. in Brain and Cognitive Sciences,
Cognitive Science track,
University of Amsterdam

supervised by
Lukas Snoek

co-assessed by
Dr. Steven Scholte

credits
12

handed in on
15th August, 2017

Contents

List of Figures

Abbreviations

ANNs Artificial neural network

CNNs Convolutional neural network

fMRI Functional magnetic resonance imaging

MPNs McCulloch-Pitts neuron

PFC Prefrontal cortex

RDMs Representational dissimilarity matrix

RNNs Recurrent neural networks

Artificial neural networks as models of information processing in biological neural networks

A.S. Wolters[1], L. Snoek[2]

[1] Master of Science 'Brain and Cognitive Sciences', Institute of Interdisciplinary Sciences, University of Amsterdam, Amsterdam, Netherlands

[2] Brain & Cognition Group, Department of Psychology, University of Amsterdam, Amsterdam, Netherlands

1 Abstract

Artificial neural networks (ANNs) are a set of computational models that were inspired by the principle of distributed processing as observed in biological neural circuits (Bailer-Jones & Bailer-Jones, 2002; McCulloch & Pitts, 1943). They are commonly-used instruments in machine learning today and have exhibited superior performance on complex tasks, most notably visual object recognition (He et al., 2016). One would assume a close link between the study of human intelligence and the engineering of intelligent systems; advances in each of the fields have, however, rarely impacted the other (Cox & Dean, 2014). Only recently have researchers started to assess whether corresponding behaviours can be observed in the mammalian visual cortex and respective ANN algorithms and have generally found comparable characteristics (Güçlü & van Gerven, 2015), prompting an investigation of ANNs as models of mammalian neural information processing (Bosch et al., 2016).

In this review, we discuss the usage of ANNs as models of information processing in biological neural networks and outline three potential benefits of using ANNs as models: firstly, such models may allow researchers to generate testable hypotheses to guide empirical investigation. Secondly, ANNs can be used to examine hypotheses that would otherwise require impractical or unethical methodologies (Izhikevich & Edelman, 2008). Thirdly, although the mechanisms that generate functions in ANN models are not readily-interpretable (Kay & Weiner, 2017), it is conceivable that further study will elucidate the mechanisms of inner workings of ANNs and, with that, provide descriptive accounts of candidate information processing mechanisms. In this review, we formulate an iterative process of successive constraining of models, establishing of correspondence between the model and target, and generation of testable hypotheses. Concrete examples of previous studies are presented.

2 Introduction

Humans have been driven to build intelligent, human-like, systems for millennia; the history of self-operating machines, so-called automata, goes as far back as the Hellenistic period (Brett, 1954). These machines were often grounded in the predominant technology of a given era. Advances in mechanic systems in general, and clockworks in specific, led to the arrival of automata with more complex structure and function, such as a humanoid robot produced by Leonardo da Vinci in 1495 (Moran, 2006). It thus comes as no surprise that the arrival of digital computing in 1946 led to a multitude of attempts to synthesise intelligent

behaviours digitally (McCarthy et al., 2006). Much was known at that time about how neural circuits process information: the neuron doctrine was commonly accepted, describing the notion that the nervous system is made up of separate cells that share no physical contact with each other (Ramón y Cajal, 1894). It had also been known that nerve cells transmit information through propagation of electrical currents, in form of spikes. These get elicited in an all-or-nothing paradigm, meaning that supra-threshold stimulation is required for a neuron to transmit a spike (Gasser & Erlanger, 1922).

A combination of other scientific advances had, however, to occur first for the idea of simulating a brain-like structure in an artificial computational substrate to be deemed feasible. Turing formulated his theory of computation, one of whose tenets stated that all computations can be described in a digital form (Turing, 1937). Advances in neuropsychology elucidated candidate mechanisms that could potentially explain learning in neural circuits (Hebb, 1949); advances in information theory allowed to further understand the nature of digital signals and their transmission (Shannon, 1948). The first practical advances in this field emanate from the University of Illinois, where Warren Sturgis McCulloch, a neurophysiologist, and Walter Pitts, a logician, published a mathematical formulation of the behaviour of a single nerve cell. They abstracted neuronal functioning as receiving one or more inputs and transforming this input into an output through an activation function. In the initial publication, this activation function took on the form of a step-wise function, which represents that an output signal is sent only if the summed input crosses a defined threshold, *i.e.* the output takes on an all-or-nothing characteristic. Their work was motivated by the view that an abstraction, or simplified description, of biophysical phenomena aids understanding and enables further inspection (McCulloch & Pitts, 1943).

Attempts have been made shortly thereafter to describe network architectures made up of multiple such nodes (Farley & Clark, 1954; Rochester et al., 1956). The first description of a neuronal network capable of learning was published by Frank Rosenblatt, a researcher of the Cornell Aeronautical Laboratory, in 1957. He described a particular arrangement of interconnected McCulloch-Pitts neurons (MPNs) which contains a layer of MPNs that feed an input forward to an output layer that also consists of MPNs; all nodes are connected to each other along the processing pathway. Frank Rosenblatt‘s main contribution was the implementation of a learning rule, *i.e.* the weights between neurons were alterable. The error, or the difference between output and target, was used to adjust the connection strengths between the nodes (Rosenblatt, 1957). The perceptron, as it was called, demonstrated the ability to learn tasks requiring logical computation to be solved; research on it was then, however, mostly dropped after Marvin Minsky and Seymour Papert published their book *“Perceptrons”* in 1969. This book discussed many of the shortcomings of perceptrons, most notably their inability to solve problems that are not linearly separable; if a classification problem is linearly separable there exists at least one straight line that separates all members of each class when these are arranged geometrically (Elizondo, 2006). Minsky and Papert argued that, to successfully solve problems that are not linearly separable, architectures consisting of multiple layers are required. The perceptron‘s learning rule can, however, not be used to train weights across multiple layers (Minsky & Papert, 1969). Interest of the scientific community to further develop ANNs declined considerably after *“Perceptrons”* was published; the following period is hence often referred to as an artificial intelligence winter (Buchanan, 2005).

A rule for efficiently propagating error signals across multiple layers to train all weights

had been formulated as early as 1974 in a Ph.D. thesis (Werbos, 1974), but the relevant algorithm was only published eight years later (Werbos, 1982). It was, however, only after David Rumelhart, Geoffrey Hinton and Ronald Williams published their formulation of the same rule that the backpropagation algorithm became widely used (Rumelhart et al., 1986). Backpropagation paved the way for further developments, as it was shown shortly thereafter that a multi-layer ANN can — theoretically — approximate any static function, *i.e.* they are universal function approximators (Hornik et al., 1989). The most notable further developments included network architectures that were directly inspired by neuroscientific findings. An example is the neocognitron (Fukushima et al., 1983) which featured a network architecture that was constrained by earlier findings of David Hubel and Torsten Wiesel describing anatomical and functional characteristics of the visual system in cats, such as localised hierarchical connections and the tuning of nodes to specific features (Hubel & Wiesel, 1962).

Despite such promising advances the computing power needed to train ANNs in reasonable time periods was lacking at the time, hence the resulting capabilities did not meet the expectations and much of the attention from the research community shifted to other, non-network-based solutions, such as support vector machines (Kriegeskorte, 2015; Vapnik & Lerner, 1963). It was only during the last decade that ANNs have received widespread attention once again. This is mostly due to the continuous acceleration of computational power, in specific through leveraging the parallel processing capabilities of graphics processing units, to more efficiently train ANNs (Oh & Jung, 2004). Equally, data sets have emerged that contain large amounts of labelled data, such as the ImageNet data set (Jia Deng et al., 2009); this enables training ANNs capable of solving ever more impressive supervised learning problems (see 3.2), *i.e.* problems where the network has access to the correct solutions. A further development of the neocognitron, the convolutional neural networks (CNNs), have recently even surpassed human-level performance in visual object recognition on still images (He et al., 2016), a highly-complex task.

In essence, a little less than 120 years after Santiago Ramón y Cajal found that the brain‘s capabilities arise from a fragmented computational substrate (1894), it has become possible to solve complex tasks with networks based on the same fundamental idea; ANNs have reached performance on these complex tasks that rivals humans‘ abilities, albeit only in specific domains (Lake et al., 2015). One would assume that a close link occurred between the study of human intelligence and the engineering of intelligent systems. However, whilst the field of ANN was initially inspired by neuroscientific knowledge, only little inspiration was drawn from biological information processing principles thereafter (Bailer-Jones & Bailer-Jones, 2002), the neocognitron being a notable exception (Fukushima et al., 1983). Equally, advances in our understanding of computational capabilities of ANNs has had little to no impact on the study of mechanisms underlying information processing in biological neural circuits (Kietzmann et al., 2017).

Recently, though, it was shown that there are striking similarities in characteristic behaviours of ANN nodes when compared to recordings of mammalian visual cortices, both recorded during visual object recognition (Cadieu et al., 2014; Güçlü & van Gerven, 2014; Khaligh-Razavi & Kriegeskorte, 2014; Yamins & DiCarlo, 2016), having led to the argument that ANNs are worthy to be examined as models of mammalian cognitive capabilities (Bosch et al., 2016; Kietzmann et al., 2017; Scholte et al., 2017). Above and beyond, ANNs can be interpreted as a homogeneous neuron-like functional substrate that allows for any capability

to arise (Hornik et al., 1989; Schäfer & Zimmermann, 2007), provided that appropriate data sets and cost function are available (O'Reilly & Frank, 2006). ANNs as functional models of cognitive functions can also be interpreted as models that inherently model multiple levels of description (see Marr, 1982, Poggio, 2012 and Lisman, 2015 for more information on levels of description in neuroscience) as they, when simulated, provide behavioural outcomes grounded in a mechanistic neural implementation (O'Reilly & Frank, 2006), making them intriguing tools for the endeavour of cognitive neuroscience (Kaplan, 2015).

In this review, we follow the notion that ANNs, when approached as models, are to be seen as a representation of the studied neural capability for the purpose of further understanding its underlying mechanisms (Frigg & Hartmann, 2016). Model-based simulations allow for manipulation in ways that are often impractical or unethical in biological organisms (Izhikevich & Edelman, 2008), meaning that a model itself can, and should, be the target of systematic study. Formulating models is an accepted scientific practice and has a long and momentous history, with the formulation of widely-known models such as the double helix model of deoxyribonucleic acid (Watson & Crick, 1953) or the Bohr model of the atom (Bohr, 1913) representing hallmark achievements within the natural sciences. Neuroscience, equally, has relied on both mathematical models and *"model organisms"* (Ellenbroek & Youn, 2016, p. 1), most often rodents, to attempt to expand the range of possible experimentation.

The statement that ANNs are good models of human functions, has, however, also come into contention. Common criticisms aim at the complexity of ANNs and state that our inability to describe them makes them unsuitable as models of cognitive functions (Kay & Weiner, 2017). ANNs have often been described as black boxes, meaning that their inner workings are not readily-understandable (Benitez et al., 1997). Equally, it has been argued that generic ANNs, with no specific structural specification, are unlikely to be able to explain the functions that arise from the brain's intricate anatomical structure (Edelman, 2015). This review aims to convince the reader of a contrary notion, *i.e.* the idea that ANNs are capable modelling tools of neural computation that will be beneficial for the endeavour of understanding the mechanisms that underlie how behavioural capabilities arise from neural circuits.

This review will start with a brief outline of the general concepts of ANNs in section three. In section four we describe, in general terms, how using ANN models can be insightful and devise an idealised process of how ANN-based models should be used in conjunction with empirical approaches; in brief, we propose that a model should allow to predict *known* experimental data (to establish correspondence between the model and the target phenomenon) but also make further, *excess* predictions that constitute testable hypotheses. Examples of previous studies will be given in sections five and six; in section five, we examine how a fully-trained network can be used to model a cognitive function by looking at CNNs and their ability to carry out object recognition on still images. In this section, we argue that previous studies have shown that CNNs fit known experimental results well, but only few attempts have been made to rigorously understand their inner workings to generate testable predictions; potential approaches are then outlined. In section six we examine an example of how training-induced changes in ANNs can model adaptation in biological neural networks. In section seven the issue of quantitatively establishing correspondence is discussed and future approaches to strengthen the relation between ANN models and their targets are described. The review is then summarised and concluded in section eight. The reader

should note that this review neither attempts to be a practical tutorial for how to use ANN machinery in the context of the neuroscientific endeavour, nor to discuss the mathematical concepts behind these networks in-depth. The aim is rather to describe the intuitions that best describe common ANNs and to analyse how this machinery can potentially lead to new insights in the neuroscientific domain.

3 Artificial neural networks: a brief overview

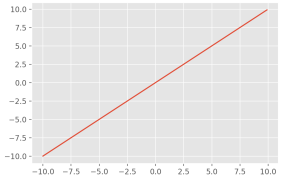
In brief, ANNs are a set of computational models that are based on an analogy, with the aim to transfer *"the idea of parallel distributed processing, as found in the brain, to the computer"* (Bailer-Jones & Bailer-Jones, 2002, p. 2). Whilst principles of neural computation informed the fundamental processing principles of ANNs, it has now become one of the major computational approaches to analyse data sets in a myriad of ways, often with no importance placed on its neurobiological inspirations and the biological plausibility of the operations that are involved (Cull, 2005). In this section, the construct of an ANN and its major components will be introduced. Many parameters can be altered to change the behaviour of ANNs; these changes are usually carried out in two different domains, in *(a)* the architectural arrangement of a network and *(b)* through alteration of the training paradigm that describes the process for optimising the network's parameters (Yamins & DiCarlo, 2016). In this section, the choices to be made in each of these areas — network architecture and learning rule — are introduced.

3.1 Architectural choices

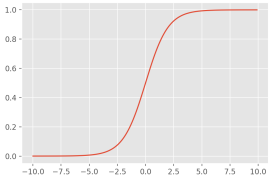
Setting up an ANN entails having to define the architecture of the model used. Choices must be made regarding *(a)* the type of processing units, or nodes, *(b)* the connection patterns between these nodes, as well as *(c)* the number of layers and number of nodes in each of those layers. To keep the distinction clear, in this review 'nodes' will be used to describe the elements of ANNs and 'neurons' will be used to describe the nervous cells as part of biological neural networks.

3.1.1 Nodes and activation functions

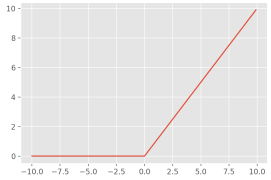
A node receives an input and transforms it into an output signal according to a defined rule. Output signals of biological neurons are spikes (Gasser & Erlanger, 1922); most ANNs, however, use static nodes to avoid heavy computational loads. This part of the review hence focusses on static nodes, spiking nodes are briefly discussed in section 7.1.3. The rules that transform the input of a node to its output are called activation functions; many different activation functions have been formulated, a discussion of which exceeds the scope of this review. Three activation functions, chosen due to their widespread usage (Karlik & Olgac, 2010), will be outlined here. Firstly, a linear activation function entails that the output of a node is equal to the summation of all its inputs and is hence also described as the identity function (see figure 1a; Rojas, 1996). A sigmoid function is a non-linear activation function that follows the logistic curve with a minimum output value approaching zero



(a) The linear function, or identity function.



(b) The sigmoid function



(c) The rectified linear unit function

Figure 1: Visualisations of three common activation function, with the x values describing a node's input and the y values describing its output. It should be noted that the scale of the y axis varies across the three graphs.

and a maximum output value approaching one (see figure 1b). It was formulated as it is differentiable, which is required for using backpropagation (Hecht-Nielsen, 1989). A rectified linear unit is a non-linear activation function that outputs zero for all inputs up to a certain threshold (often defined at point zero); if the summation of the inputs crosses the threshold its linear summation will be transmitted (see figure 1c; Dahl et al., 2013).

3.1.2 Connection patterns

There are two main aspects of connection patterns that need to be defined, firstly the pattern of connections and, with that, the information flow that is implemented, which vastly impacts the capabilities of the given ANN (Moody, 1994). Secondly, the selectivity of connections needs to be defined, *i.e.* whether all nodes are connected to all other nodes in the successive layer, or if selective connection patterns are to be implemented.

Neural circuits in mammalian brains display three types of connections: feedforward (feeding information onto neurons in the successive layer), lateral (feeding information onto neurons in the same layer) and feedback (feeding information onto neurons in the previous layer; Rojas, 1996). As mentioned, this impacts the flow of information and, with that, crucially alters a network's computational capabilities. The vast majority of ANN models only feature feedforward connections as robust and efficient training regimes have not yet been formulated for networks featuring more complex connection patterns (Pascanu et al., 2012). If a network contains feedback connections it is a network with bi-directional, or recurrent, information flow characteristics, which are called recurrent neural networks (RNNs). RNNs are noteworthy as they can represent an additional dimension, often described as the ability to represent time, or act based on contextual information (Botvinick & Plaut, 2006); this network type will be briefly discussed in section 7.1.2.

CNNs are decent examples to illustrate what the concept of selective connectivity entails. These networks models were first constructed in an attempt to constrain a generic ANN to more closely match the architectural organisation that had been observed in the visual processing pathway of mammals, including the phenomenon that neurons would only respond to stimuli in a small patch of the input space, called its *receptive field* (Hubel & Wiesel, 1962). This selective connectivity, *i.e.* that a certain column of a network only processes

a patch of the input space, was replicated in CNNs by selectively connecting nodes within so-called *kernels* (Lecun et al., 1998).

3.1.3 Depth and width of the network

Networks can be endowed with varying numbers of layers and nodes in each of these. Networks are described as 'deep' when they contain more than one layer that is neither receiving external inputs nor expressing the network's output (Deng & Yu, 2014). Networks with a vast number of layers have become commonplace after Alex Krizhevsky demonstrated superior performance in visual object recognition on the ImageNet challenge with a network made up of eight such hidden layers (Krizhevsky et al., 2012). Extreme numbers have also recently been showcased by Microsoft Research, members of which published a paper that describes a network with 152 layers in total (He et al., 2016). The choice of building deeper networks, often with less nodes in each layer, has been backed up theoretically by researchers of the Weizmann Institute of Science, who brought forward a proof that depth *"can be exponentially more valuable than width"* (Eldan & Shamir, 2015, p. 1). It has been hypothesised that increasing the depth of a network is akin to enabling the network to decompose the task at hand in incrementally finer functional fractions (O'Reilly & Frank, 2006). It is noteworthy, however, that excessively-deep feedforward ANNs are not biologically plausible, as humans are capable of recognising objects after as little as 150 milliseconds, which, if the comparably slow processing speeds of neurons are considered, does not warrant the involvement of excessively-deep hierarchical structures in the sensory processing of mammalian brains (Thorpe et al., 1996).

3.2 Choice in learning rules

The most capable ANN models were trained using error-driven training regimes, hence these algorithms will be the focus of this introductory section. More biologically-inspired learning rules, such as Hebbian learning, will be discussed in section 7.1.1. ANN training regimes are usually described as either supervised, meaning that the network has access to the correct solutions during training, or unsupervised, meaning that no solutions are given (Schmidhuber, 2015); other forms such as semi-supervised (Hady & Schwenker, 2013) and reinforcement learning (Stanley & Miikkulainen, 2002) also exist. This review will focus on supervised training regimes, also due to their widespread usage. In a supervised training regime for a classification problem, training is carried out based on a data set that contains true labels of what class each example belongs to. Crucially, the network has access to these labels; usually, some data points are used to evaluate the network's efficiency in achieving a certain task, hence labels are removed from selected data points, a process that is called cross-validation (Bishop, 1994).

Supervised training regimes contain three main components, a cost function, a rule for the propagation of errors and an update rule. The cost function, or loss function, defines the calculation of the overall error; this is crucial for supervised training regimes as the objective is to minimise that loss function. This minimisation of the cost is achieved by updates of the connection weights between nodes as defined by the update rule. Knowing the overall error in the network's output is not sufficient to know how to update specific

connection weights; to achieve this the errors must be propagated through the network and each weight's impact on the error needs to be defined. This is described by the error propagation rule (Schmidhuber, 2015).

The most widely-used training regime is backpropagation, which has been hugely influential and requires stochastic gradient descent, which describes a method to update weights based on a calculation of the gradient of the loss function, unveiling the direction in which a weight update would improve performance on the given training example (Rumelhart et al., 1986). It follows from this that no formal definition of the precise computation to be carried out is given during training of ANNs. Hence, as a solution to a particular problem is not defined explicitly, but rather reached by the algorithm itself, ANNs have often been referred to as 'black boxes' (Benitez et al., 1997). More details on learning rules and their biological plausibility will be given in 7.1.1.

This review does not attempt to describe all of the many ANN models that have been presented; it is impossible to do so in the given scope. There are many specific types of ANNs, such as echo state networks, Boltzmann machines, long short-term memory model or self-organising maps, only to name a few; for a more complete introduction to the types of ANNs, please refer to the comprehensive introduction by Jürgen Schmidhuber (2015).

4 Models in science

Building models is of fundamental importance to science. This review follows the definition of a scientific model as an abstract representation of a certain target phenomenon that is being studied (Frigg & Hartmann, 2016). In the context of ANN models of neural phenomena it can be argued that every model implements a candidate version of the actual mechanism that generates the phenomenon in question (Kietzmann et al., 2017), hence ANNs should be understood as instances of generative models (Arakaki et al., 2017; Edelman, 2015). In this review, we also follow the notion that the models are an instantiation of all tenets of a scientific theory; *i.e.* all generative models correspond to a theory about a potential mechanism of the target. This section will outline three advantages that the use of models could entail. Firstly, we will briefly describe an iterative process that allows models to more closely correspond to what is known about their targets whilst simultaneously providing empirical researchers with testable hypotheses. This general process is based on dual prediction; models need to allow for prediction of *known* experimental data to establish correspondence, but equally predict additional data to guide further experimentation (*excess prediction*; see section 4.1). Secondly, models can be used to carry out experiments that are either impractical or unethical to be carried out in biological organisms; hence many yet untested hypothesis are potentially testable (see section 4.2). Thirdly, provided that correspondence has been quantitatively established, a model represents a potential mechanism that is assumed to be causing the target phenomenon and a descriptive account of a model's inner workings can hence lead to the formulation of new explanatory accounts (see section 4.3). Examples of these approaches will be briefly introduced alongside, but discussed in later sections.

4.1 A process of correspondence and hypotheses

The potential importance of model predictions is best described by examining an example from a different scientific endeavour, physical cosmology. The existence of dark matter is widely accepted in this scientific community; no experimental evidence has, however, been gathered to prove its existence. Dark matter is hypothetical, its existence was formulated to explain unsolved observations. With that, dark matter is a prediction from the standard model of cosmology (Bertone et al., 2004) and the existence of dark matter remains a hypothesis to be tested. We see one of the main benefits of model building in the ability to generate testable hypotheses. Hereafter we devise an idealised iterative process for integrating empirical experimentation and model creation.

(Step 1) A new model needs to be generated by incorporating known characteristics of the target systems, a process that is known as constraining. To give an example, one of the network architectures that is most widely used for visual object recognition (the CNN) was formulated to closely match what was then known about the mammalian visual system, as briefly discussed in 3.2. Constraining was based on neuroanatomical and functional evidence; firstly, it was known that the human visual system was hierarchically structured, *i.e.* that cells in earlier layers are involved in the processing of simple features, whereas cells in the later layers are responsive to more complex features (Hubel & Wiesel, 1962). It was also known that specific cell clusters in layers of the mammalian visual system are only connected to a specific subset of cells in the later layers (Hubel & Wiesel, 1962); this connection pattern has been replicated, first in the neocognitron (Fukushima et al., 1983), then in the CNN (Krizhevsky et al., 2012). Constraining is further discussed in 7.2.

(Step 2) Correspondence between the model and its target phenomenon needs to be established. As a first step, implausibilities must be alleviated; these occur when a model implements a mechanism (or part thereof) in a way in which it cannot possibly be implemented in the target system. Secondly, known experimental data from the target phenomenon should be compared to data that ANN models are able to predict, *e.g.* characteristics of node activations such as tuning or task performance; we will refer to this prediction as *known prediction* hereafter. It should be noted that matching predictions must not be interpreted as evidence that the same mechanism is in place in both the model and the target, but rather more vaguely as 'something similar is likely occurring'. A variety of other approaches to quantify correspondence have been devised; these will be described in section 5.1.

(Step 3) If correspondence can be successfully established, this equates to having created a well-informed candidate mechanism that potentially constitutes the target function. To further assess the target *excess predictions* must be made. This often requires systematic study of the model's behaviour (examples are given in the following sections) as well as a mapping to the expected data, *e.g.* mapping from an observed pattern of node activation in the ANN model to an expected spiking sequence in respective biological circuits. Excess predictions ideally take on the form of empirically-testable hypotheses.

(Step 4) Empirical research could then falsify or accept these hypotheses. If a hypothesis can be confirmed by empirical studies, the respective *excess prediction* can then also be interpreted as a *known prediction*, further strengthening the correspondence of the model to the target.

(Step 5 and following) This process is assumed to be iterative. Falsified hypotheses require updates to the model, confirmed hypotheses can usually be followed up by other, more granular hypotheses until a full descriptive account of the constituting mechanisms has been provided.

This describes a very general process that is constituted by phases of model generation, establishing correspondence, hypothesis generation and empirical hypothesis testing. It must be mentioned that this process is idealised; only few studies have approached ANN-based modelling with such rigour. A notable example comes from the field of predictive coding theory which hypothesises that the brain continuously predicts sensory inputs through its top-down connections, whereas bottom-up connections carry the error of said predictions (a more detailed discussion follows in section seven; Clark, 2013). Modelling was carried out with recurrent neural networks that successfully predicted known data about receptive field properties (*known prediction*); based on this model it was then hypothesised that the source populations of the top-down and bottom-up signals must be segregated (*excess prediction*; Rao & Ballard, 1999), a hypothesis that was confirmed in later studies in mice (Berezovskii et al., 2011) and macaques (Markov et al., 2014).

4.2 Models to test hypotheses

As was mentioned briefly before, it has been argued that models allow to be systematically studied in ways not commonly available in biological organisms (Izhikevich & Edelman, 2008); in the context of ANN models this would entail the systematic disruption of aspects of a neural circuit to assess its effects on network functions. Systematic disruption of neural functioning in humans is only available within the limits of what is considered ethical and through techniques such as transcranial magnetic stimulation Walsh & Cowey (2000). It is, however, important to note that this approach requires to quantitatively establish, or at least explicitly assume, a two-fold correspondence; not just the model has to correspond to the system in question, it is also required to establish a correspondence between the manipulation of the model and the manipulation in the target system that the hypothesis is drawn from. To give an example, a recent study tested whether a change in cost functions would lead to hypothesised effects on the parameter overlap in networks (Scholte et al., 2017); cost function manipulation here is seen as a model of hierarchical decomposition of behavioural goals. This study is described more thoroughly in section six.

4.3 Descriptive accounts of model‘s mechanisms

The immense importance of models in the history of science has been laid out earlier in this review; it is noteworthy, though, that the Bohr model of the atom (Bohr, 1913) and

the double helix model of deoxyribonucleic acid (Watson & Crick, 1953) not only provide predictions but also a comprehensive descriptive account of the phenomenon in question. ANNs as models, however, are not descriptive models as such; they should first and foremost be viewed as simulations of a potential mechanism (Gerstner et al., 2012) that enable us to make predictions. If a model is capable to make predictions that match experimental data we can interpret that a candidate mechanisms has been *captured* by the model, which does, however, *not* entail that we automatically know more about the mechanism in question (Gao & Ganguli, 2015). In essence, modelling a human function with an ANN substitutes one barely-understood system with another; this is an often-formulated criticism against using ANNs as models at all (Kay & Weiner, 2017). Based on the assumption that both systems are, in theory, comprehensible, we should argue that we replace one system that is impractical and often unethical to study with another one that allows access to all its parameters and enables any conceivable manipulation procedure, as is the case with ANNs (Yamins & DiCarlo, 2016). This constitutes completeness of data recordings and manipulative powers that are unlikely to be available in biological circuits in the foreseeable future (Gao & Ganguli, 2015; Izhikevich & Edelman, 2008). ANNs have been described as 'black boxes' as the precise mechanism of solving a certain task is not defined *a priori* (Benitez et al., 1997); it should, however, not be inferred from this that understanding the inner workings of ANNs is unachievable.

The differences between ANN models and their respective target systems are often emphasised to argue against the use of ANNs as models; the employed nodes are often static rather than spiking (Kay & Weiner, 2017), mammalian visual cortices feature widespread feedback connections, that have recently been shown to be a main driver behind visual cortex activity, but are not commonly implemented in ANNs (Markov et al., 2014) and the major training regimes are unlikely to be biologically plausible (Bengio et al., 2015), only to mention a few of the common simplifications. What appears more striking to us, however, is the conceptual similarity that exists between the capabilities of CNNs and object recognition in the human cortex as they share one crucial characteristic with biological networks: the ability to solve highly complex tasks emerges from the interplay of large numbers of computationally-simple entities. We argue here that the ability to adequately describe how function emerges from the interplay of such computationally-simple entities is crucial for formulating adequate descriptions of both artificial and biological neural networks. ANNs, with all their parameters being accessible, are hence ideal models to understand the methodological requirements for formulating complete mechanistic descriptions. To our knowledge, no such complete description of the mechanisms behind this emergent capability has been brought forward in ANNs.

ANN models can be studied in two ways, either through mathematical theory, the *theoretical approach*, or through simulation, which constitutes the *experimental approach* (Gerstner et al., 2012). A theory of deep learning, derived from mathematical descriptions of the functions that occur, is still in its infancy. Advances have, however, recently been made with regards to how adding layers allows for more complex functions to be computed (Bianchini & Scarselli, 2014), that saddle points dominate the learning dynamics (Dauphin et al., 2014) and how input statistics are represented through synaptic changes as a result of learning (Saxe et al., 2013). Approaches to understanding ANN functioning through an analysis of its simulations, *i.e.* the experimental approach, are outlined in section 5.2. Whilst understanding the inner workings of ANNs is undoubtedly important on its own, our ability to describe it has further implications for methods in neuroscience. Applying experimental

approaches to ANNs allows us to understand just how impactful a certain methodological approach can potentially be, as we consider the ideal case with all parameters of a neural systems being recordable. Equally, assuming the hypothetical case that the theoretical approach has led to insightful descriptions, then these descriptions firstly constitutes testable hypotheses; secondly, as it is known what parameters were needed to be derive these descriptions, statements can be made with regards to what type of data collection needs to be collected in biological organisms.

One should infer from this that studying how function emerges within ANNs is imperative for reasons other than understanding mammalian functions directly; it is equally important from a methodological standpoint. If we are unable to adequately describe how function emerges from the interplay of nodes in ANNs by employing methodology usually employed in neuroscience, it appears rather unlikely that the very same methods will allow us to formulate mechanistic accounts of how functional capabilities arise from the interplay of neuronal firing, a central aim of the field of cognitive neuroscience (Bechtel, 2008). Attempting to sufficiently describe the mechanisms underlying functional emergence in ANNs is a crucial proving ground to test and further develop neuroscientific methodologies as well as discuss the appropriateness of different descriptions (Eliasmith, 2010); it has been argued that this would allow to address crucial methodological concerns such as the following: *"Even if we could collect any kind of detailed measurements about neural structure and function, what theoretical and data analytic procedures would we use to extract conceptual understanding from such measurements?"* (Gao & Ganguli, 2015, p. 1). A recent paper is noteworthy in this context; researchers attempted to describe the mechanistic functioning of a microprocessor, which allows for any kind of detailed measurements, with a battery of many commonly-used neuroscientific methods and ultimately failed to do so (Jonas & Kording, 2017). This opinion can also be seen as an argument against the statement that relevant mechanistic understanding will be elicited purely by the availability of more detailed neural recordings (Lloret-Villas et al., 2016) and simulations (Markram et al., 2011); we highly doubt that the mechanisms will become clear to us by simply generating more data. Rather we will be crucial for gaining an understanding of how bigger data sets should be analysed (Gao & Ganguli, 2015).

In summary, we believe that ANNs, as models of neural information processing in biological circuits, can be beneficial by generating testable hypotheses, providing new means for testing hypotheses and to provide description of candidate mechanisms. Equally, assessing which methodological approaches allow to adequately describe ANN mechanisms is an important scenario to further develop neuroscientific methods. The next sections describe concrete examples of how ANNs have previously been used as models of neural information processing.

5 Trained artificial neural networks as models of specific capabilities

In this section, the use of trained ANNs as models of neural information processing will be discussed. Whilst ANNs have been applied to a variety of tasks, they have arguably had the most impact in the field of object recognition due to the superior performance of CNNs

(He et al., 2016; Krizhevsky et al., 2012). We will start this section by outlining the notions behind CNNs; this description will be followed by subsections on establishing correspondence and methods that could potentially allow to generate experimentally-testable hypotheses.

CNNs, as previously outlined, are a type of feedforward ANNs with architectural parameters set to resemble characteristics reminiscent to those of the mammalian visual system. CNNs are a further development of the neocognitron (Fukushima et al., 1983). To give a simplified sketch of its workings, CNNs contain three layer types: convolutional layers, pooling layers and fully-connected layers. Nodes are not connected to all nodes in the successive layers, but rather to a certain subset. In convolutional layers, each node represents a filter that is convolved over the section of the input volume that the selectively-connected subset is tuned to; abstractly, these layers detect features in the input image by applying filters to each image position. Pooling layers reduce the dimensionality; intuitively this entails that having identified a certain feature is deemed to be computationally more important than retaining its exact location, *i.e.* feature representations become invariant to the location in which the original features occurred. Fully-connected layers then form the output of the network, which, in the context of object recognition, is often a vector representing the likelihood that the image contains an instance of each class of objects that the network knows about; the highest likelihood value represents the object the network has recognised in the input (Krizhevsky et al., 2012). In the next section, 5.1, methods for establishing correspondence will be outlined; section 5.2 then introduces methods of directly studying the behaviour of ANNs.

5.1 Establishing correspondence between CNNs and visual cortex recordings

As outlined in section four, correspondence needs to be established, usually through prediction of some previously-collected experimental data which we refer to as *known prediction*. Performance levels have been shown to be comparable between CNNs and humans (He et al., 2016). Furthermore, the tendency of nodes to respond to gradually more complex stimuli, as observed in the human visual system (Hubel & Wiesel, 1962), has also been observed in CNNs (Zeiler & Fergus, 2013); these response profiles are, however, difficult to compare. To address these difficulties in establishing correspondence two methodological approaches, representational similarity analysis and encoding models, have recently been devised and are described in sections 5.1.1 and 5.1.2, respectively.

5.1.1 Representational similarity analysis

Representational similarity analysis entails to compare response similarities between sets of measures. It is based on the idea that, as we cannot reliably compare representations in biological and artificial networks directly, we should compare these in a different space. Representational geometry assumes that a system‘s representation of different stimuli forms a multidimensional space, with each dimension representing one point of measurement, *e.g.* recorded feature activity of an artificial node, a recorded functional magnetic resonance imaging (fMRI) voxel or a recorded single-neuron response profile. This forms a space that describes all possible representations of a given neural system (Kriegeskorte & Kievit, 2013).

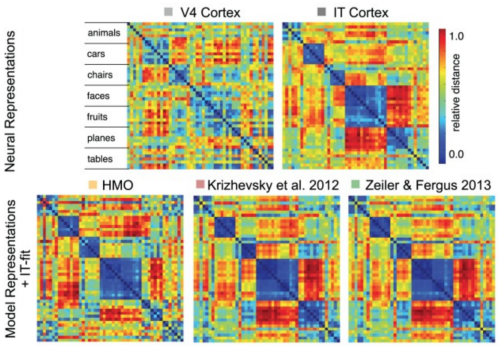
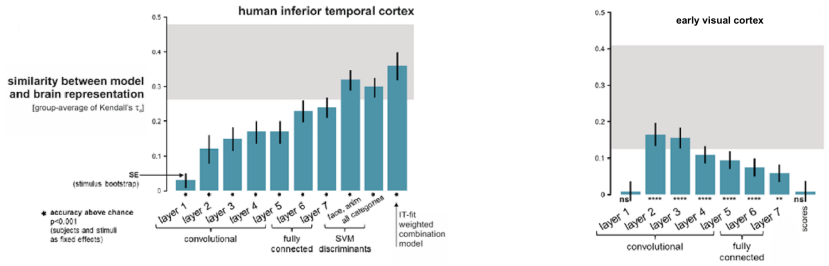


Figure 2: RDMs of human cortical recordings (first row) and three neural network models (second row); HMO stands for hierarchical modular optimization algorithm (Yamins et al., 2014). Images are replicated along the x and y scales; the diagonal line hence describes the comparison of the same images viewed which explains why the representational differences are approaching zero values. Taken from Cadieu et al., 2014.

Pairs of stimuli are analysed to compare the representational geometry between the human visual system and a CNN. For each of the image pairs the dissimilarity in the respective activation patterns is calculated; representational dissimilarity matrices (RDMs; see figure 2) can be drawn from this. To describe the representational space that occurs in the human visual cortices neural recordings are analysed, commonly collected via fMRI (Kriegeskorte et al., 2008) or electroencephalography (Kaneshiro et al., 2015). These were collected whilst human subjects were actively viewing different images. Feature activities are drawn from a forward-pass of the two respective images in a CNN. The resulting RDMs are then compared with a rank-based correlation measure (*e.g.* Spearman's r, Kendall's τ; Khaligh-Razavi & Kriegeskorte, 2014). Representational similarity analysis can hence be described as an attempt to quantify the correspondence between feature spaces by abstracting from the input space (Kriegeskorte, 2015); it analyses how closely the representation profiles match between two systems. Example of RDMs are shown in figure 2.

A few studies have used representational similarity analysis to understand the correspondence between CNNs trained for object recognition and the mammalian visual system. Firstly, it was attempted to understand if the characteristic increasing complexity in representations from early to late visual areas of mammalian visual cortices also occurs in CNNs. The RDMs of all CNN layers were compared to data from early human visual areas and the human inferior temporal cortex, which is thought to underlie the higher-order object recognition capabilities in humans (Lehky & Tanaka, 2016); RDMs of early CNN layers showed close resemblance with those derived from lower visual human areas whereas RDMs of later CNN layers showed closer correspondence to the ones derived from human inferior temporal cortex, as measured with Kendall's τ (Khaligh-Razavi & Kriegeskorte, 2014); see figure 3.

A second analysis assessed different object recognition algorithms and found that the structural organisation of the measured representations in the algorithm resembles this of the human inferior temporal cortex more closely with increasing object recognition performance (Cadieu et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014); whilst



(a) (b)

Figure 3: Kendall's τ to calculate the similarity between the RDMs derived from all CNN layers and *(a)* the human inferior temporal cortex and *(b)* the early human visual cortex in areas V1, V2 and V3. Taken from Khaligh-Razavi & Kriegeskorte, 2014.

it cannot be inferred from this that an algorithm must model the brain's visual system closely to achieve high performance, the indicator is nonetheless striking (see figure 4). *Vice versa*, it can also be inferred that ANN-based computer vision models that show superior performance are more likely to be able to explain representational data from the human inferior temporal cortex via the representational similarity analysis well (Khaligh-Razavi & Kriegeskorte, 2014; Kriegeskorte, 2015).

5.1.2 Encoding models

Encoding models provide another option to establish correspondence by using the detected features of an ANN to predict neural responses. In an encoding model a CNN is trained for object recognition, but the feature activations that occur during a given task are used as an input to a separate response model which is trained to predict neural activity; this prediction is then compared to the observed neural response via a Pearson correlation coefficient r (van Gerven, 2016). This allows researchers to achieve two things: firstly to establish correspondence, secondly to provide a further analysis of the neural correlates. If an encoding model reaches high predictive accuracies, it can be argued that the model encapsulates the information that is available in the given set of neural recordings (van Gerven, 2016). A decoding model can be derived from an encoding model, which is able to carry out the reverse prediction, *i.e.* from neural data to the stimulus (Naselaris et al., 2011). This allows researchers to understand the informational content of neural recordings even further; decoding accuracies can be compared across brain regions which allows to infer the amount of information about the stimulus that is retained by the recorded brain areas (Horikawa et al., 2013). Decoding is further discussed in 5.2.2.

In a relevant study that attempted to establish correspondence between the human visual system and CNNs, fMRI recordings were collected of human participants actively viewing images. Predictions were made based on the feature activity of all layers and it was found that earlier layers of the CNN better predicted the neural response of earlier visual areas, with later CNN layers predicting activity in later human visual areas with higher accuracies (Güçlü & van Gerven, 2015), see figure 5.

Recent work has also shown that encoding models are not just applicable to responses

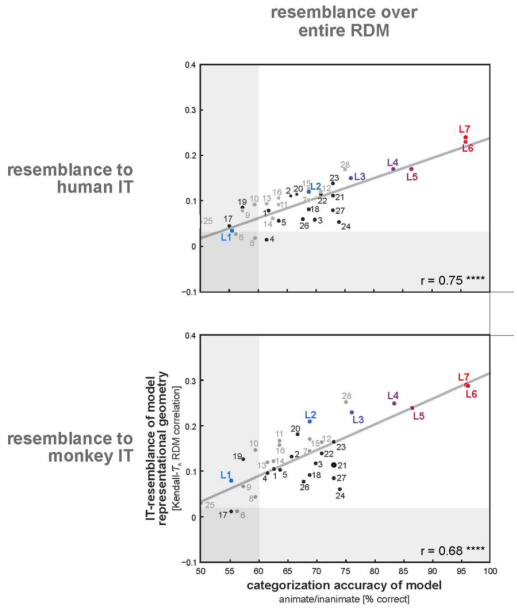


Figure 4: This figure plots the performance of a variety of object recognition systems along their performance (x axis) and their representational resemblance with the human and monkey inferior temporal cortex (y axis); a strong correlation was found. Taken from Khaligh-Razavi & Kriegeskorte, 2014.

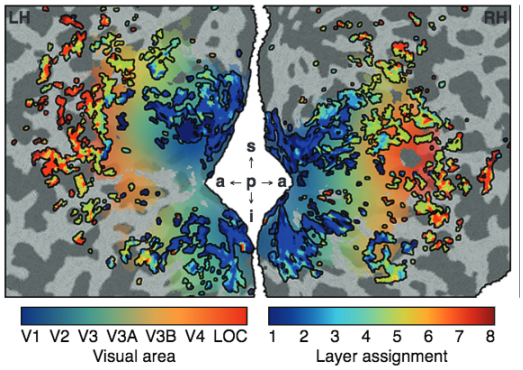


Figure 5: This graph visualises how encoding models can be beneficial for the analysis of functional anatomic data. The less-saturated tint shows which areas a voxel is defined as, anatomically; the overlaid, stronger-saturated colours show which CNN layer best predicted the neural response of a given area. Taken from Güçlü & van Gerven, 2015.

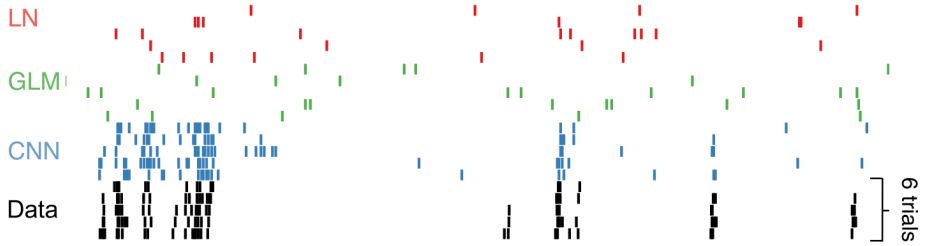


Figure 6: This graph shows spike predictions from convolutional neural network to neural signals observed in the frog's retina. The last row depicts the spikes extracted from neural recordings from a frog's retina, the row above it is the prediction from a CNN. LN stands for linear-nonlinear model, GLM for generalised linear model. Taken from McIntosh et al., 2017.

to static images, but that sequenced retinal responses of a frog — representing the very first transfer from light stimuli to neural signals — can be modelled efficiently with a deep CNN and a coupled response model; the correlation between the CNN-based encoding model and the retinal responses was shown to be 0.64, which meant that its accuracy in prediction significantly exceed previous results from linear-nonlinear models and generalised linear models (McIntosh et al., 2017); see figure 6.

Equally, it has been tested whether RNNs can be used to extract more optimal features as these networks are able to retain information about the stimulus history through their recurrent dynamics. In this first study it has been shown that RNNs, in combination with a response model, can reasonably capture the neural response to a sequence of images, as measured by fMRI (Güçlü & van Gerven, 2016); more evidence is, however, required to fully understand the nature of features generated by RNNs, see also section 7.1.2.

5.2 Experimental approaches to CNN functioning

In the previous section we outlined a number of studies that found a reasonable correspondence between CNNs and the human visual system. This has led a variety of researchers to conclude that the examination of CNNs as models of such perceptual processing is warranted (Bosch et al., 2016; Kietzmann et al., 2017; Scholte et al., 2017). As stated in section four, CNNs can be thought of as complete implementations of candidate mechanisms for visual object recognition and should hence be used to provide testable hypotheses through simulation and/or theoretic analysis. There is, however, a notable lack of such testable predictions made on the basis of an assessment of CNNs, which is often attributed to our inability to understand the mechanisms underlying the functioning of such networks (Benitez et al., 1997). Attempting to improve our understanding of ANNs has been argued for in section four. In summary, this approach is valuable for two reasons: a better understanding of the mechanisms underlying ANNs would allow us to *(a)* generate appropriate, testable predictions and *(b)* to provide descriptive accounts of what potential mechanism each ANN implementation represents.

In the following part of this review we outline a variety of such experimental approaches.

In the first subsection, virtual neurophysiology, a field of study entailing the experimental manipulation of ANNs or their inputs, is introduced. In the following subsection we analyse whether methods attempting to understand neural population coding in biological neural networks are applicable to ANNs.

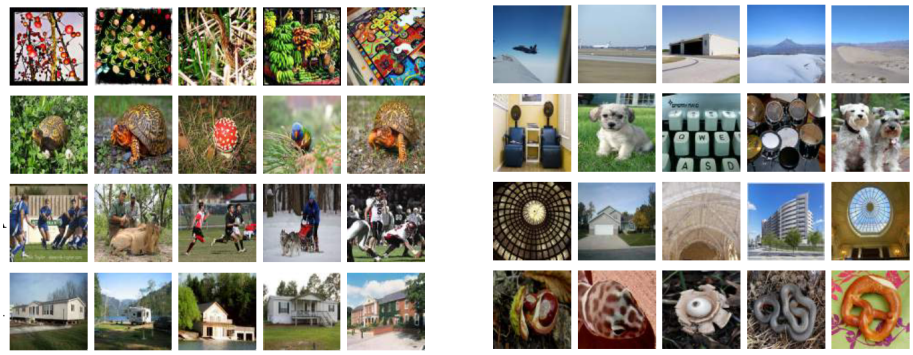
5.2.1 Virtual neurophysiology

Neurophysiology is a field of study that attends to the functioning of the neural system through spatio-temporal recordings of the activity of its components in correlation to some task (Carpenter & Reddi, 2012). In the first subsection, we will outline methodological approaches that systematically alter parts of the network through perturbation; the latter subsection will describe attempts at systematic manipulation of network inputs.

Perturbing artificial neural networks In a recent paper, Dan Yamins and James DiCarlo sketched out the idea of what they call virtual electrophysiology, aiming to establish a methodological framework that entails systematic disruption of network structures to understand the effects this causes. The authors state that this approach might potentially allow for causal inferences to be drawn; if a perturbation of a node, or a cluster of nodes, leads to a predicted behavioural alteration, then this finding should be deemed as evidence for a causal link between the phenomena that occur on these two levels of description (Yamins & DiCarlo, 2016). In general, all parameters of ANNs are accessible and changeable; *i.e.* perturbations or lesions can be carried out with relative ease. Previous studies are, unfortunately, rare; the most relevant previous study attempted to test the assumptions underlying double dissociations, an experimental approach in psychology that states that, if a lesion leads to selective disruption in one cognitive function but not the other and *vice versa*, then these functions are thought to be caused by entirely-independent mechanisms (Teuber & Hans-Lukas, 1955). This was simulated in a study from 1993, in which a simple ANN model was trained to carry out two separate mappings from input to output whilst nodes were selectively inactivated; the authors found that double dissociations can be observed through selective manipulation of nodes in ANNs (Bullinaria & Chater, 1993) and hence provided evidence in favour of the validity of double dissociations as a method.

Experimentally examining tuning of nodes As outlined above, it has been observed that nodes in CNNs during the task of recognising objects show similarities to the behaviour of biological neurons; nodes only activate when a certain feature or combination of features appears in the input data. Nodes hence exhibit a behaviour reminiscent of feature tuning in biological neurons (Güçlü & van Gerven, 2015). A few studies have attempted to further examine which nodes are responsive to which inputs.

Firstly, it has been attempted to visualise the input that is likely to lead to the highest activation in a single neuron - or a population of neurons - to examine their tuning. Agrawal *et al.* used a CNN to predict what input would most *(1)* activate or *(2)* deactivate neurons from each of four areas of the visual pathway (V1, V4, extrastriate body area and the parahippocampal place area). The mapping was established by understanding which layers of the CNN best predicted the neural activity in each area (see the section on encoding



(a) (b)

Figure 7: These are the images that most *(a)* activate, or *(b)* deactivate a node in the CNN layer that best encodes a brain area; rows one to four represent V1, V4, extrastriate body area and the parahippocampal place area respectively. The lower layers most activate to the occurrence of patterned images; the later layers most activate to the occurrence of separated objects. Taken from Agrawal et al., 2014.

models in 5.1.2). This offers a qualitative impression of what stimuli neurons in these areas are tuned to (Agrawal et al., 2014); see figure 7.

Secondly, representations of what a node is tuned to can be gained through reversing the analysis, *i.e.* through adding another convolutional network with the same, but reversed architecture, that fully reverses the process of convolution (Zeiler & Fergus, 2013). This deconvolution approach generates intriguing insights into the tuning of nodes, but also offers only qualitative results; see figure 8 for examples.

Another insightful approach is called prediction difference analysis, which attempts to visualise if a certain pixel was used as evidence for, or against, the chosen network output. The measure was derived by extending a paradigm for estimating relevance values of features (Robnik-Sikonja & Kononenko, 2008). A comparison of AlexNet, VGG16 and GoogLeNet, all commonly-used ANN algorithms, showed that very different strategies seem to be occurring despite the fact that all networks correctly classified the objects in the images; see examples in figure 9. It should be noted, however, that the methodology did not include an attempt to quantify the dissimilarity between the network architectures (Zintgraf et al., 2016) and hence also only offers qualitative evidence.

5.2.2 Population coding in artificial neural networks

The notion that a focus not on single node behaviour but rather an assessment of the interactions within clusters of nodes is likely to lead to new insights is relatively new to the neuroscientific community (Yuste, 2015a). A recent study found that, in a trained ANN, information is encoded in the space of activations, with perturbations of single units rarely influencing the network‘s behaviour (Szegedy et al., 2013). This supports the proposed notion that clusters of neurons lead to emergent properties and activation patterns of single nodes do not carry much identifiable information (Yuste, 2015a), a finding that is related

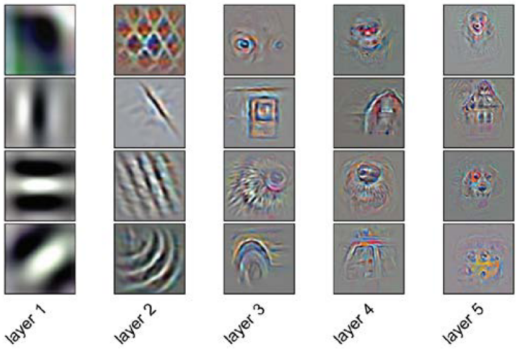


Figure 8: Examples of deconvolution across different layers. Lower layers are made up of simple shapes, whereas higher layers show more complex combinations of shapes; taken from Kriegeskorte, 2015.

(a) (b)

Figure 9: Prediction difference analaysis was carried out on two images that were analysed by either the AlexNet, the GoogLeNet or the VGG16 network; red pixels represent evidence for the chosen class, blue pixels against, grey pixels were disregarded. Fundamental differences in the problem solving strategies between networks appear to occur. Taken from Zintgraf et al., 2016.

to the concept of redundancy in biological neural circuits which describes the fact that a disruption of the functioning of a single neuron usually does not disrupt the overall function of a network (Schneidman et al., 2003). This finding can also be related to the framework of population coding, in which it is argued that responses of single neurons represent a stimulus-dependent distribution which result in noisy measures; coding these responses entails an estimation of the likely stimulus. It is also argued that, for the purpose of coding, the response patterns of a single neuron is relatively uninformative (Averbeck et al., 2006). Addressing questions related to population coding in humans usually entails using one of two experimental paradigms, namely decoding and Shannon information theory (Quiroga & Panzeri, 2009); hereafter we will outline each of these and review studies that attempted to apply these methodologies to ANNs.

Decoding approaches to population coding Decoding was briefly mentioned in section 5.1.2 and describes algorithms that are capable of predicting a dimension of the stimulus from neural data. The technique is called 'reconstruction' if the stimulus is reproduced in its entirety (Naselaris et al., 2011). Fundamentally, the brain itself encodes and decodes information as part of its processing; firstly, sensory inputs are encoded into neural signals which are then constantly transformed and decoded to be made actionable (Bialek et al., 1991; Eliasmith & Anderson, 2003). This has generally been conceptualised as upstream neurons linearly reading out the information that is encoded in downstream neurons (Kriegeskorte & Kievit, 2013).

Decoding paradigms allow researchers to understand the information content in different pathways by attempting to output the stimulus that was presented, or a dimension of it (Naselaris et al., 2011). This is an established approach to understand the information content that neuronal populations code for and has helped to give evidence for the encoding of object and person identities in neurons of the human medial temporal lobe. A study found that the type of person or object could be decoded near-perfectly from a population recording in this area; specific images of a certain object or person could, however, not be decoded at all (Quiroga et al., 2007). Whilst ANNs have been used as tools to improve the decoding and reconstructing of stimuli from human neural recordings (Naselaris et al., 2011), no study has attempted, to the best of our knowledge, to use decoding on ANNs directly. It could be hypothesised that stimuli can be decoded/reconstructed with differing accuracies along the processing hierarchy, which would be an intriguing result. This endeavour might be worthwhile as it is known that up to 90% of the node activities of a CNN during object recognition take on a zero value (Foroosh et al., 2015). As a result, one must assume that sub-networks of the CNN carry out the object recognition. Identifying these might allow us to research phenomena reminiscent to neural reuse in humans, which describes that neural circuits often carry out more than one function (Anderson, 2010). This will be further discussed in the next section.

Shannon information theory approaches to population coding Information theory is an area of mathematics that studies how information can be quantified, stored and distributed; it was initially proposed as a formalisation of the study of signal processing (Shannon, 1948). Information theory as a description of information flow has recently gained popularity in neuroscience as a tool to describe the flow of information in biological neural networks (Lungarella et al., 2006). One of the key information theoretic measures

is entropy, which quantifies the amount of uncertainty of a process (Kullback, 2008). In 2000, a measure of statistical coherence between system, called transfer entropy, was formulated (Schreiber, 2000); it is assumed that this measure allows researchers to quantify the information flow between neurons and neuronal clusters (Overbey & Todd, 2009). In an application of this measure to fMRI data, a recent study in participants with traumatic brain injury has found that transfer entropy measures were reliable at describing the precise impacts of these injuries on functional connectivity (Mäki-Marttunen et al., 2013). It is important to note that functional connectivity here describes the estimated information flow between neuronal populations and is hence a very different characterisation to anatomical, or structural, connectivity. Structural connectivity describes the pathways that information can possibly flow along, effective connectivity describes the pathways that are actually used to solve a task (Ito et al., 2011).

Interestingly, transfer entropy as a measure of effective connectivity has been applied to the study of ANNs as well. In a recent study an evolutionary algorithm was used to create a number of spiking neural networks with varying parameters. These networks then acted as controllers for virtual agents whose task was to move towards one type of object and move away from the other, a type of visual classification task. The transfer entropy measures for each trial, representing the information flow between nodes, were then clustered. Two main clusters formed, representing one condition each. The authors conclude that a given network presents itself as different networks depending on the task requirement through vastly varying information flows. It was also concluded that networks that show strong within-task homogeneity and across-task heterogeneity are likely to show better task performance (Vasu & Izquierdo, 2017). These findings are intriguing in the light of neural reuse theories (Anderson, 2010) and the large number of zero values that commonly occurs in CNNs (Foroosh et al., 2015). As briefly mentioned in the section about decoding, analysing the information flow, or the ability to decode over certain areas, might allow a description of functional networks, *i.e.* those parts of the network that effectively carry out a task.

6 Training-induced changes in artificial neural networks as models of adaptation

In the previous section, we have outlined that fully-trained CNNs can act as models of visual object recognition, albeit more research is needed to generate testable hypotheses. In this section, we propose that learning-induced changes of ANNs can be informative models of adaptation in biological neural networks. As previously mentioned, this review focusses on supervised training regimes. In this section, we will briefly review a recent study that tested the hypothesis that decomposition of cost functions, in an assumed correspondence to the compartmentalisation of a main behavioural goal into constituting subgoals, leads to the emergence of functionally-specific neuron clusters.

Even early neuroscientific evidence described cognitive functions as localised; this conclusion was reached through studies on patients with brain lesions that led to very specific impairments, most famously when Paul Broca observed that localised lesions led to characteristic speech impairments (Broca, 1861). The two-stream hypothesis of vision, one example

of such localised functional specification, describes a clear distinction between two higher visual processing pathways: the dorsal stream is thought to be responsible for visually-guided behaviours and the ventral stream is thought to underlie object recognition (Goodale & Milner, 1992). There have been a variety of attempts to understand the driving forces behind this functional dichotomy of visual processing. An interesting recent model describes this dichotomy as driven by a decomposition of cost functions (Scholte et al., 2017), a perspective that is grounded in a recent conceptualisation that views the underlying driving force of much of the functional specificity of brain areas as a result of different cost functions being optimised in each of the areas (Marblestone et al., 2016).

In this study, an ANN was trained with two separate tasks conditions; it was then tested whether training the network on these two related tasks leads to a different functional connectivity pattern than using two unrelated tasks. The hypothesis was that training a network on two qualitatively different tasks will lead to a functional circuit split, akin to the two visual processing streams. What they found was that, when the ANN was trained on two unrelated tasks, most nodes tended to be involved with only one task; this was not the case when the same network was trained on two related tasks, in line with their initial hypothesis (Scholte et al., 2017). They also derived a method to quantify this sharing of feature representations across multiple tasks in a single network; it can be seen to be somewhat analogous to the information theoretic measure of transfer entropy that was outlined in section 5.2.2. This study is an interesting example of hypothesis testing in models, *i.e.* the second potential benefit of using models outlined in section four.

7 Correspondence, biological plausibility and future research

Much of this review has described potential benefits of using ANNs as models of information processing in biological neural circuits, in line with recent opinions (Bosch et al., 2016; Kietzmann et al., 2017; Scholte et al., 2017). It is necessary, however, to further discuss the issue of establishing correspondence between the model and its target. In essence, the issue with using ANNs as models of human functioning is that *"an explicit correspondence between model parameters and physical variables is missing"* (Mohamad & Reza, 2016, p. 5). All areas of science that leverage models to illustrate or explain specific phenomena must assume a strict correspondence between the model and the target system it models (Frigg & Hartmann, 2016). This assumption is difficult to ascertain in the context of ANNs and neural circuits for a number of reasons. Firstly, ANNs are vastly simplified abstractions of their target systems, *i.e.* biological neural circuits. Many aspects of ANNs are not plausible in biological neural circuits; an intriguing example is the phenomenon of universal adversarial perturbations. Research has shown that visual object recognition in CNNs can be manipulated through systematic alteration of the input image; these changes cause the CNN to misclassify the image but remain imperceptible to a human observer (Moosavi-Dezfooli et al., 2016). In summary, it has to be argued that the correspondence between biological neural systems and ANNs is *"far from fully understood"* (Yamins & DiCarlo, 2016, p. 363).

This issue even remains with representational similarity analysis. Let us assume, hypothetically, that we find a set of RDMs that show a perfect match between the biological and

artificial recordings; this does, however, not allow us to infer that the same representation is learnt in both systems. The structure of the respective RDMs may be identical, but it remains conceivable that these are driven by different features, *e.g.* ANN responses that are principally driven by luminance when the brain representations were driven by another set of principal features, representing the external world differently. In other words, establishing correspondence using representational similarity analysis is an ill-posed problem.

In line with the process laid out in section four, we propose a three-fold research effort to approach correspondence; firstly, biologically implausible characteristics of current ANN models must be alleviated and the biological realism of such models must be strengthened. Secondly, ANN models need to be constrained directly by data collected from neural recordings. Thirdly, the constrained models need to be used to make known predictions about the target systems; these predictions can then be tested to probe the validity of the model. In this section we will further discuss the alleviation of biological implausibilities (section 7.1) and constraining of models (section 7.2); approaches to hypothesis generation have been outlined in section 5.2.

7.1 Improving biological realism

ANNs are vastly simplified models of neural circuits; alleviating biological implausibilities is hence crucial for improving the correspondence to biological circuits. In this section, we will put three examples forward; firstly, the biological implausibility of many learning rules (see 7.1.1), the addition of recurrent information processing (see 7.1.2) and the use of spiking nodes (see 7.1.3).

7.1.1 On the biological plausibility of learning rules

Whilst a variety of algorithms have been brought forward that approach the problem of learning in supervised or unsupervised contexts, versions of backpropagation of error algorithms (see 3.2) are the most widely used class of supervised learning algorithms (Schmidhuber, 2015) that have also led to many of the most impressive results on highly complex tasks (He et al., 2016). It has, however, been argued that this class of algorithms is not biologically plausible, *i.e.* that backpropagation cannot possibly occur during learning in biological circuits. This coincides with the observation that biological inspirations have been drawn from intensely to develop well-suited neural architectures; only minimal inspirations have, however, been drawn from neuroscience in the development of training paradigms and learning rules (Wadhwa & Madhow, 2016), although first examples of capable learning algorithm derived from neural processing principles have recently been presented (Chandrashekar & Granger, 2012).

Using a biologically implausible algorithm hinders the correspondence between ANN-based models and the biological targets of the model, especially if learning-induced changes are to be tested. Ideally, ANNs would not be probed as models if they are grounded in aspects that are clearly incompatible to what is known in neuroscience (Frigg & Hartmann, 2016); it can be argued, however, that the representations that result from backpropagation regimes are biologically plausible, which is sufficient to warrant the validity of using these ANNs as models. Equally, requiring a plausible learning rule would entail that ANNs with

the ability good performance on complex functions, like human-level object recognition, are unobtainable with current plausible training regimes, significantly reducing the appeal of ANNs as models.

It has to be argued that there is a fundamental divide between the learning rules commonly used to train ANNs and the principles of learning as observed in biological neural circuits. ANNs have acquired the ability to solve complex computation by relying on external error, or *ground truth*, signals to compute the error that is propagated backwards through the networks; this signal is not available to a learning organism in most situations. Error signals are, however, commonplace in biological neural circuits; the question of how these are generated has received some attention recently (den Ouden et al., 2012). It has been argued that organisms acquire error signals through prediction that is then compared to the perceived outcome, which has been backed up by studies showing that dopaminergic neurons code for reward prediction errors (Schultz, 2016). Equally, neurons have been observed that specifically code for errors that occur when predictable sensory stimuli are not observed (Keller et al., 2012). This framework, usually described as predictive coding, has the potential to integrate the error-based learning paradigm of supervised learning that is often successfully used in machine learning with the associative, or Hebbian, learning paradigm that appears to be occurring in the brain; first such algorithms have been proposed, *e.g.* in a research paper that showed how local associative learning processes might suffice to approximate the credit assignment that occurs during backpropagation processes (Bengio et al., 2017); predictive coding is also discussed in this review within the context of understanding recurrent processing (see 7.1.2). It is hence imperative to explore biologically-plausible training regimes although progress is hindered by the fact that principles of biological learning are not yet fully understood (Bengio et al., 2015). Whilst alleviating implausibilities is certainly a step in the right direction, further testing and constraining would be necessary; this would require abstracting neural data into models, or constraining these via available recordings (Frigg & Hartmann, 2016).

7.1.2 Recurrent network architectures

The brain displays vast lateral and feedback connectivity (Kriegeskorte, 2015) and is hence best described as an RNN. RNNs are defined as a class of ANNs which feature circular connections, *i.e.* the models also contain feedback connections. In theoretical terms, this creates an internal state, or reservoir, within the network that acts akin to a memory state and allows the networks to act upon contextual or sequential information (Pearlmutter, 1989). There is, however, another property of RNNs that makes them beneficial for modelling cognitive functions. As previously noted, purely-feedforward ANNs are universal approximators of static functions (Hornik et al., 1989); RNNs, however, are universal approximators of dynamical systems (Schäfer & Zimmermann, 2007), a property that allows extraction of mathematical descriptions of how state variables change over time (Sussillo & Barak, 2013).

Recurrence as a modelling paradigm Whilst the usage of RNNs for more complex tasks is still in its infancy due to the difficulty of efficiently training these networks (Pascanu et al., 2012), intriguing first results have recently been brought forward for the case of having to recognise objects that are partly occluded by other objects. Feedforward CNNs

have displayed significant difficulties with this task, with adequate performance having been achieved only recently, and only when objects are easily-segmentable (Chandler & Mingolla, 2016). A comparative study has examined whether the addition of *(1)* only lateral connections or *(2)* lateral and feedback connections improves the performance; it was shown that, under light occlusion, only the addition of lateral connections improved the performance. The addition of feedback connections did, however, significantly improve performance when objects were heavily occluded (Spoerer et al., 2017); whether this improves correspondence measures to recordings of mammalian visual cortices has not yet been assessed.

The computational role of lateral and feedback connections, which are found in most cortical areas (Markov et al., 2014), is far for from clear. One of the most influential theoretical explanations of the role of this recurrence is predictive coding, which views the brain akin to a hierarchical predictive machine that constantly generates hypotheses, or predictions, about sensory inputs and updates these in the face of contradicting evidence (Thornton, 2017; Clark, 2013), as briefly introduced in section 7.1.1. In this framework the top-down connections would carry information about the expected sensory input; the forward connection would then carry the signals of whether this input has been matched by sensory inputs (Rao & Ballard, 1999). There are a number of models of predictive coding, many of which are based on ANN-like structures (Spratling, 2015), that provide predictions and direction for further research. One such model-derived prediction stated that the source populations between the feedback connections (carrying predictions) and the feedforward connections (carrying the prediction error) must be entirely separate (Rao & Ballard, 1999), which was found empirically only recently (Berezovskii et al., 2011; Markov et al., 2014).

Understanding and modelling the human brain‘s recurrence likely involves approaches that integrate a more holistic view on the relation between perception and action than the commonly use mapping of stimulus to response (Edelman, 2015). A recent study has backed this statement up; the researcher recorded neurons of the early visual cortex in mice *in-vivo* and, surprisingly, found that *“visual input alone is a poor predictor of activity in primary visual cortex”* (Keller et al., 2012, p. 810), leading to the conclusion that, in behaving animals, visual cortex activity cannot be explained without understanding the role of motion-related signalling. First attempts at devising experimental paradigms are being made by using a version of reinforcement learning algorithms that allow ANNs to be trained to carry out extremely complicated tasks in the context of simulated agents in virtual reality environments (Bosch et al., 2016).

Extracting dynamical systems descriptions from RNN models Dynamical systems theory is a branch of mathematics that leverages differential and/or difference equations to understand the behaviour of complex systems over time (Luenberger, 1979); due to the non-linear dynamics of its components, *i.e.* the neurophysiological characteristics of neurons, the brain can be viewed as such a complex dynamic system and approached with the same tools (McKenna et al., 1994). Analysing the brain as such a dynamical system entails to study the change of variables in form of their trajectories in a state space. It can be stated that dynamical systems analysis offers a separate, abstract level of description that characterises the behaviour of a system in terms of *“higher-level variables describing global states of the system”* (Kaplan & Craver, 2011, p. 604).

Trajectories of state variables can be reverse-engineered from RNNs with relative ease

through finding fixed or slow points in the state space (Sussillo & Barak, 2013); this following section will describe a recent study of decision-making in the macaque prefrontal cortex (PFC) of macaques and how RNN modelling as well as dynamical systems analysis tools helped to analyse the data in a meaningful way. In this study, macaques were trained to carry out a sensorimotor task that involved making context-dependent decisions based on task-related and task-unrelated visual information. An RNN was trained in an analogous task; the temporal dynamics profile and trajectories were then reverse-engineered from this RNN. It was found that context-dependent decision-making can be described fully by the interaction of two components, the line attractor and the selection vector. The line attractor, which is a mechanism to acquire evidence for, or against, a decision (Sussillo, 2014), remained stable in the state space. The selection vector, however, shifted its position depending on the task demands, delivering a description of a possible mechanism for the selective integration of task-relevant information, although it does not entail a formulation of neural computation. To establish correspondence, the firing rates of the RNN model were compared to the neural recordings of the PFC in macaque monkeys and resulted in a reasonable fit (Mante et al., 2013).

It is currently being debated whether understanding the dynamic properties of a neural circuit can be understood as explaining how function emerges. There appears to be an agreement that formulations describing a cognitive function in dynamical systems terms do not fully elucidate the mechanism in questions as (neural) implementational details are lacking (Kaplan & Craver, 2011); whether such a model can be deemed explanatory is being debated (Ross, 2015). Furthermore, it has been argued that models formulated in the dynamic systems theory framework are *"extremely difficult to confirm or refute in light of the vast majority of neural data"* (Eliasmith, 2010). It is, however, undoubtedly the case that dynamical system models of neural functions possess descriptive powers and, through harnessing a different perspective, allow the research community to generate interesting descriptive accounts and testable hypotheses.

7.1.3 Spiking nodes

Biological plausibility remains a concern on the detailed implementational level of description when ANNs are used as models of biological functions; most, if not all, large-scale ANN models employ vastly simplified, non-spiking processing units. Biological neurons elicit spiking outputs, and ensembles often synchronise in rhythmic oscillations, which is a phenomenon that is thought to be crucial for establishing efficient communication between neuronal ensembles (Masuda & Aihara, 2002). Whilst the exact computational importance of spikes is still being discussed (Tiesinga et al., 2008), it is worth noting that the population coding methods outlined in section 5.2.2 are used to understand recordings that capture spiking data. It follows from this that spiking nodes are required for many of these analyses to function; equally, using spiking nodes would increase the biological plausibility of the model in place. Networks containing spiking nodes have indeed been formulated and shown to be trainable (Diehl & Cook, 2015; Mesnard et al., 2016); state-of-the-art object recognition performance has also recently been presented (Hunsberger & Eliasmith, 2016). These training regimes were, however, rate-based, meaning that they are using the number of spikes over a defined period as an input. A single-spike-based supervised learning rule has been published recently (Zenke & Ganguli, 2017); further research is, however, needed

to test its applicability and reliability.

7.2 Constraining models with empirical data

Constraining models by precise neurophysiological measures has been difficult even in the recent past due to our inability to record from large numbers of neurons *in-vivo* (Yuste, 2015b). Small-scale functional data recording methods have been vastly improved in recent years, with the conveyance of data from hundreds of neurons simultaneously now becoming feasible with two-photon calcium imaging (Hamel et al., 2015). This progress is unlikely to be halted in near future, as recent funding initiatives are aimed at improving these imaging techniques, *e.g.* through the National Institutes of Health's BRAIN initiative (Insel et al., 2013). Hence, suitable data is likely to be increasingly available in the near future, at least when non-human recordings are considered.

The question of how to best constrain an ANN to match experimental data has, unfortunately, not been discussed much in the scientific literature. ANNs were, as briefly mention before, often constrained to match what is known only about neural architectures (Wadhwa & Madhow, 2016), as is the case in the neocognitron (Fukushima et al., 1983). One recent study is, however, noteworthy as it was attempted to train an ANN to match response profiles directly derived from neurophysiological data. Their goal was to generate a model of the centre-surround suppression effect as observed in the primary visual cortex. The network, an RNN, was trained with backpropagation; the cost function that was optimised described the difference between near-surround behaviour as observed in experimental data and the behaviour of ANN nodes. Intriguingly, it was found that the network behaviour after training not only corresponded to the training data but also generalised, *i.e.* that other known behaviours could be observed that were not part of the training set (Lotfi et al., 2014). Whilst this first study serves as little more than a feasibility study - only simulated experimental data was used and no functional capabilities emerged from the network - we do believe that this is potentially a beneficial area for further research. It would be intriguing to understand if, and how, ANNs can be trained to both match experimental recordings, as a matter of constraining, and acquire some functional capabilities, to generate hypotheses.

8 Conclusion

In summary, as ANNs are forms of distributed processing algorithms that were inspired by the way the brain functions (Bailer-Jones & Bailer-Jones, 2002), they can be leveraged to model and simulate brain functions. CNNs have been shown to be well-corresponding models of object recognition as it is occurring within mammalian visual pathways (Cadieu et al., 2014; Güçlü & van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014); learning processes of ANN models have been used to test whether the manipulation of cost functions can potentially explain the modularisation of cortical function and, with that, the development of the anatomic duality that is observed within visual processing (Scholte et al., 2017). RNNs have showed promise for modelling even more complex tasks than object recognition (Spoerer et al., 2017); relevant progress in training ANNs based on precise spike-timing has been presented very recently (Zenke & Ganguli, 2017), which is of importance for improving the biological plausibility of ANN models.

There are, however, issues in establishing and quantifying the correspondence between ANNs and the targets they model (Yamins & DiCarlo, 2016); we believe that more rigorous approaches to ANN modelling, described as an iterative process of constraining, prediction and hypothesis testing, can potentially alleviate this problem. We strongly encourage the further study of ANNs, experimentally and theoretically, as reaching a better understanding of the inner workings of ANNs might enable us to *(a)* draw informed inferences on the capabilities of the respective methods and *(b)* generate more well-informed hypotheses to be empirically tested. To conclude, ANNs as models are certainly still in their infancy; we do believe, though, that the unique combination of these models to allow for systematic study and manipulation (Izhikevich & Edelman, 2008), their reliance on the same fundamental computational structure as the brain (Kriegeskorte, 2015) and their high performance on increasingly complex tasks (He et al., 2016; Spoerer et al., 2017) are some of the reasons why ANNs are likely to take on a major role within the neuroscientific endeavour in future.

9 Acknowledgements

I would like to thank Lukas Snoek for his superb supervision of this literature review; I highly appreciate our interesting discussions and the vast amount of insight that I have gained in the process of writing this review. Equally I would like to thank Dr. Steven Scholte for making this thesis possible and his co-assessment in the process. Furthermore, I would like to thank Stephan Garbin for the immensely useful inputs from the perspective of a machine learning researcher; this has taught me that interdisciplinarity is key to any endeavour. I also want to thank Francesca Mazzone and Giorgio Manenti for the fruitful inputs and interesting discussions.

Word count 11,767 / 12,000

References

Agrawal, P., Stansbury, D., Malik, J., Gallant, J. (2014). Pixels to Voxels: Modeling Visual Representation in the Human Brain. *arXiv*, pages 1–15.

Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and Brain Sciences*, 33(04):245–266.

Arakaki, T., Barello, G., Ahmadian, Y. (2017). Capturing the diversity of biological tuning curves using generative adversarial networks. *arXiv*.

Averbeck, B. B., Latham, P. E., Pouget, A. (2006). Neural correlations, population coding and computation. *Nature Reviews Neuroscience*, 7(5):358–366.

Bailer-Jones, D. M. Bailer-Jones, C. A. (2002). Modelling data: Analogies in neural networks, simulated annealing and genetic algorithms. *Model-based reasoning: science, technology, values*, page 147.

Bechtel, W. (2008). *Mental mechanisms : philosophical perspectives on cognitive neuroscience*. Routledge.

Bengio, Y., Lee, D.-H., Bornschein, J., Mesnard, T., Lin, Z. (2015). Towards Biologically Plausible Deep Learning. *arXiv*.

Bengio, Y., Mesnard, T., Fischer, A., Zhang, S., Wu, Y. (2017). STDP-Compatible Approximation of Backpropagation in an Energy-Based Model. *Neural Computation*, 29(3):555–577.

Benitez, J. M., Castro, J. L., Requena, I. (1997). Are artificial neural networks black boxes? *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, 8(5):1156–64.

Berezovskii, V. K., Nassi, J. J., Born, R. T. (2011). Segregation of feedforward and feedback projections in mouse visual cortex. *The Journal of Comparative Neurology*, 519(18):3672–3683.

Bertone, G., Hooper, D., Silk, J. (2004). Particle Dark Matter: Evidence, Candidates and Constraints. *arXiv*.

Bialek, W., Rieke, F., de Ruyter van Steveninck, R., Warland, D. (1991). Reading a neural code. *Science*, 252(5014):1854–1857.

Bianchini, M. Scarselli, F. (2014). On the Complexity of Neural Network Classifiers: A Comparison Between Shallow and Deep Architectures. *IEEE Transactions on Neural Networks and Learning Systems*, 25(8):1553–1565.

Bishop, C. (1994). Novelty detection and neural network validation. *IEE Proceedings - Vision, Image, and Signal Processing*, 141(4):217.

Bohr, N. (1913). On the constitution of atoms and molecules. *Philosophical Magazine Series 6*, 26(151):1–25.

Bosch, S. E., Seeliger, K., van Gerven, M. A. J. (2016). Modeling Cognitive Processes with Neural Reinforcement Learning. *bioRxiv*.

Botvinick, M. M. Plaut, D. C. (2006). Short-term memory for serial order: A recurrent neural network model. *Psychological Review*, 113(2):201–233.

Brett, G. (1954). The Automata in the Byzantine 'Throne of Solomon'. *Speculum*, 29(3):477–487.

Broca, P. (1861). Perte de la parole, ramollissement chronique et destruction partielle du lobe anterieur gauche du cerveau. *Bull Soc Anthropol*, 2(1):235–238.

Buchanan, B. G. (2005). A (Very) Brief History of Artificial Intelligence. *AI Magazine*, 26(4):53.

Bullinaria, J. Chater, N. (1993). Double dissociation in artificial neural networks: Implications for neuropsychology. *Proceedings of the fifteenth annual conference of the cognitive science society*, pages 283–288.

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., DiCarlo, J. J. (2014). Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLoS Computational Biology*, 10(12).

Carpenter, R. H. S. R. H. S. Reddi, B. (2012). *Neurophysiology: a conceptual approach*. Hodder Arnold.

Chandler, B. Mingolla, E. (2016). Mitigation of Effects of Occlusion on Object Recognition with Deep Neural Networks through Low-Level Image Completion. *Computational Intelligence and Neuroscience*, 2016:1–15.

Chandrashekar, A. Granger, R. (2012). Derivation of a novel efficient supervised learning algorithm from cortical-subcortical loops. *Frontiers in Computational Neuroscience*, 5(January):1–17.

Clark, A. (2013). Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):1–73.

Cox, D. D. Dean, T. (2014). Neural networks and neuroscience-inspired computer vision.

Cull, P. (2005). Recent Developments in Biologically Inspired Computing. *Choice: Current Reviews for Academic Libraries*, 42(7):1261.

Dahl, G. E., Sainath, T. N., Hinton, G. E. (2013). Improving deep neural networks for LVCSR using rectified linear units and dropout. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8609–8613. IEEE.

Dauphin, Y., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., Bengio, Y. (2014). Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *arXiv*, pages 1–14.

den Ouden, H. E. M., Kok, P., de Lange, F. P. (2012). How Prediction Errors Shape Perception, Attention, and Motivation. *Frontiers in Psychology*, 3:548.

Deng, L. Yu, D. (2014). Deep Learning: Methods and Applications. *Foundations and Trends in Signal Processing*, 7(3-4):197–387.

Diehl, P. U. Cook, M. (2015). Unsupervised learning of digit recognition using spike-timing-dependent plasticity. *Frontiers in Computational Neuroscience*, 9:99.

Edelman, S. (2015). The minority report: some common assumptions to reconsider in the modelling of the brain and behaviour. *Journal of Experimental & Theoretical Artificial Intelligence*, 3079(September):1–26.

Eldan, R. Shamir, O. (2015). The Power of Depth for Feedforward Neural Networks. *arXiv*, pages 1–30.

Eliasmith, C. (2010). How we ought to describe computation in the brain. *Studies in history and philosophy of science*, 41(3):313–20.

Eliasmith, C. Anderson, C. H. (2003). *Neural Engineering: Computation Representation and Dyamics in Neurobiological Systems*. MIT Press.

Elizondo, D. (2006). The Linear Separability Problem: Some Testing Methods. *IEEE Transactions on Neural Networks*, 17(2):330–344.

Ellenbroek, B. Youn, J. (2016). Rodent models in neuroscience research: is it a rat race? *Disease models & mechanisms*, 9(10):1079–1087.

Farley, B. Clark, W. (1954). Simulation of self-organizing systems by digital computer. *Transactions of the IRE Professional Group on Information Theory*, 4(4):76–84.

Foroosh, H., Tappen, M., Penksy, M. (2015). Sparse Convolutional Neural Networks. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 806–814.

Frigg, R. Hartmann, S. (2016). Models in Science. *Stanford Encyclopedia of Philosophy*, pages 1–19.

Fukushima, K., Miyake, S., Ito, T. (1983). Neocognitron: A Neural Network Model for a Mechanism of Visual Pattern Recognition. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-13(5):826–834.

Gao, P. Ganguli, S. (2015). On Simplicity and Complexity in the Brave New World of Large-Scale Neuroscience. *arXiv*.

Gasser, H. Erlanger, J. (1922). A study of the action currents of nerve with the cathode ray oscillograph. *American Journal of Physiology–*.

Gerstner, W., Sprekeler, H., Deco, G. (2012). Theory and Simulation in Neuroscience. *Science*, 338(6103).

Goodale, M. A. Milner, A. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25.

Güçlü, U. van Gerven, M. A. J. (2014). Unsupervised Feature Learning Improves Prediction of Human Brain Activity in Response to Natural Images. *PLoS Computational Biology*, 10(8).

Güçlü, U. van Gerven, M. a. J. (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *Journal of Neuroscience*, 35(27):10005–10014.

Güçlü, U. van Gerven, M. A. J. (2016). Modeling the dynamics of human brain activity with recurrent neural networks. *Frontiers in Systems Neuroscience*, pages 1–19.

Hady, M. F. A. Schwenker, F. (2013). *Semi-supervised Learning*. Springer, Berlin, Heidelberg.

Hamel, E., Grewe, B., Parker, J., Schnitzer, M. (2015). Cellular Level Brain Imaging in Behaving Mammals: An Engineering Approach. *Neuron*, 86(1):140–159.

He, K., Zhang, X., Ren, S., Sun, J. (2016). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 11-18-Dece, pages 1026–1034.

Hebb, D. O. (1949). *The organization of behavior: A Neuropsychological Approach*. John Wiley & Sons, New York, NY.

Hecht-Nielsen (1989). Theory of the backpropagation neural network. In *International Joint Conference on Neural Networks*, pages 593–605 vol.1. IEEE.

Horikawa, T., Tamaki, M., Miyawaki, Y., Kamitani, Y. (2013). Neural Decoding of Visual Imagery During Sleep. *Science*, 340(6132):639–642.

Hornik, K., Stinchcombe, M., White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366.

Hubel, D. H. Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106–154.2.

Hunsberger, E. Eliasmith, C. (2016). Training Spiking Deep Networks for Neuromorphic Hardware. *arXiv*.

Insel, T. R., Landis, S. C., Collins, F. S. (2013). The NIH BRAIN Initiative. *Science*, 340(6133).

Ito, S., Hansen, M. E., Heiland, R., Lumsdaine, A., Litke, A. M., Beggs, J. M. (2011). Extending Transfer Entropy Improves Identification of Effective Connectivity in a Spiking Cortical Network Model. *PLoS ONE*, 6(11):e27431.

Izhikevich, E. M. Edelman, G. M. (2008). Large-scale model of mammalian thalamocortical systems. *Proceedings of the National Academy of Sciences of the United States of America*, 105(9):3593–3598.

Jia Deng, Wei Dong, Socher, R., Li-Jia Li, Kai Li, Li Fei-Fei (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE.

Jonas, E. Kording, K. P. (2017). Could a Neuroscientist Understand a Microprocessor? *PLOS Computational Biology*, 13(1):e1005268.

Kaneshiro, B., Perreau Guimaraes, M., Kim, H.-S., Norcia, A. M., Suppes, P., Zoccolan, D. (2015). A Representational Similarity Analysis of the Dynamics of Object Processing Using Single-Trial EEG Classification. *PLOS ONE*, 10(8):e0135697.

Kaplan, D. M. (2015). Explanation and Levels in Cognitive Neuroscience. In *Handbook of Neuroethics*, pages 9–29. Springer Netherlands, Dordrecht.

Kaplan, D. M. Craver, C. F. (2011). The Explanatory Force of Dynamical and Mathematical Models in Neuroscience: A Mechanistic Perspective. *Philosophy of Science*, 78(4):601–627.

Karlik, B. Olgac, A. (2010). Performance analysis of various activation functions in generalized MLP architectures of neural networks. *International Journal of Artificial Intelligence and Expert Systems (IJAE)*, 1(4):111–122.

Kay, K. N. Weiner, K. S. (2017). Principles for models of neural information processing. *bioRxiv*.

Keller, G. B., Bonhoeffer, T., Hübener, M. (2012). Sensorimotor Mismatch Signals in Primary Visual Cortex of the Behaving Mouse. *Neuron*, 74(5):809–815.

Khaligh-Razavi, S. M. Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, 10(11).

Kietzmann, T. C., McClure, P., Kriegeskorte, N. (2017). Deep Neural Networks In Computational Neuroscience. *bioRxiv*.

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, 1:417–446.

Kriegeskorte, N. Kievit, R. A. (2013). Representational geometry: integrating cognition, computation, and the brain. *Trends in cognitive sciences*, 17(8):401–12.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., Bandettini, P. A. (2008). Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey. *Neuron*, 60(6):1126–1141.

Krizhevsky, A., Sutskever, I., Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances In Neural Information Processing Systems*, pages 1–9.

Kullback, S. (2008). Information theory and entropy. In *Model based inference in the life sciences: A primer on evidence*, pages 51–82. Springer New York, New York, NY.

Lake, B. M., Salakhutdinov, R., Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, 350(6266).

Lecun, Y., Bottou, L., Bengio, Y., Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lehky, S. R. Tanaka, K. (2016). Neural representation for object recognition in inferotemporal cortex. *Current Opinion in Neurobiology*, 37:23–35.

Lisman, J. (2015). The Challenge of Understanding the Brain: Where We Stand in 2015. *Neuron*, 86(4):864–882.

Lloret-Villas, A., Daudin, R., Le Novère, N. (2016). Big Data in neuroscience: open door to a more comprehensive and translational research. *Big Data Analytics*, 1(1):5.

Lotfi, E., Araabi, B. N., Ahmadabadi, M. N., Schwabe, L. (2014). Biological constrained learning of parameters in a recurrent neural network-based model of the primary visual cortex. In *2014 21th Iranian Conference on Biomedical Engineering (ICBME)*, pages 292–297. IEEE.

Luenberger, D. G. (1979). *Introduction to dynamic systems: theory, models, and applications*. Wiley.

Lungarella, M., Sporns, O., Edelman, G., Sporns, O., Edelman, G. (2006). Mapping Information Flow in Sensorimotor Networks. *PLoS Computational Biology*, 2(10):e144.

Mäki-Marttunen, V., Diez, I., Cortes, J. M., Chialvo, D. R., Villarreal, M. (2013). Disruption of transfer entropy and inter-hemispheric brain functional connectivity in patients with disorder of consciousness. *Frontiers in Neuroinformatics*, 7(24):1–11.

Mante, V., Sussillo, D., Shenoy, K. V., Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, 503(7474):78–84.

Marblestone, A. H., Wayne, G., Kording, K. P. (2016). Toward an Integration of Deep Learning and Neuroscience. *Frontiers in computational neuroscience*, 10:94.

Markov, N. T., Vezoli, J., Chameau, P., Falchier, A., Quilodran, R., Huissoud, C., Lamy, C., Misery, P., Giroud, P., Ullman, S., Barone, P., Dehay, C., Knoblauch, K., Kennedy, H. (2014). Anatomy of hierarchy: Feedforward and feedback pathways in macaque visual cortex. *Journal of Comparative Neurology*, 522(1):225–259.

Markram, H., Meier, K., Lippert, T., Grillner, S., Frackowiak, R., Dehaene, S., Knoll, A., Sompolinsky, H., Verstreken, K., DeFelipe, J., Grant, S., Changeux, J.-P., Saria, A. (2011). Introducing the Human Brain Project. *Procedia Computer Science*, 7:39–42.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information, Henry Holt and Co.* MIT Press, New York, NY.

Masuda, N. Aihara, K. (2002). Bridging Rate Coding and Temporal Spike Coding by Effect of Noise. *Physical Review Letters*, 88(24):248101.

McCarthy, J., Minsky, M. L., Rochester, N., Shannon, C. E. (2006). A proposal for the Dartmouth summer research project on artificial intelligence. *AI Magazine*, 27(4):12–14.

McCulloch, W. S. Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4):115–133.

McIntosh, L. T., Maheswaranathan, N., Nayebi, A., Ganguli, S., Baccus, S. A. (2017). Deep Learning Models of the Retinal Response to Natural Scenes. *arXiv*.

McKenna, T., McMullen, T., Shlesinger, M. (1994). The brain as a dynamic physical system. *Neuroscience*, 60(3):587–605.

Mesnard, T., Gerstner, W., Brea, J. (2016). Towards deep learning with spiking neurons in energy based models with contrastive Hebbian plasticity. *arXiv*.

Minsky, M. Papert, S. (1969). *Perceptrons.* M.I.T. Press, Cambridge, MA.

Mohamad, A. Reza, M. (2016). Computational Models in Neuroscience: How real are they? A Critical Review of Status and Suggestions. *Austin Neurology & Neuroscience*, 1(2):1–10.

Moody, J. (1994). Prediction Risk and Architecture Selection for Neural Networks. In *From Statistics to Neural Networks*, pages 147–165. Springer Berlin Heidelberg, Berlin, Heidelberg.

Moosavi-Dezfooli, S.-M., Fawzi, A., Fawzi, O., Frossard, P. (2016). Universal adversarial perturbations. *arXiv*.

Moran, M. E. (2006). The da Vinci Robot. *Journal of Endourology*, 20(12):986–990.

Naselaris, T., Kay, K. N., Nishimoto, S., Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–410.

Oh, K.-S. Jung, K. (2004). GPU implementation of neural networks. *Pattern Recognition*, 37(6):1311–1314.

O'Reilly, R. C. Frank, M. J. (2006). Making Working Memory Work: A Computational Model of Learning in the Prefrontal Cortex and Basal Ganglia. *Neural Computation*, 328:283–328.

Overbey, L. Todd, M. (2009). Dynamic system change detection using a modification of the transfer entropy. *Journal of Sound and Vibration*, 322(1):438–453.

Pascanu, R., Mikolov, T., Bengio, Y. (2012). On the difficulty of training Recurrent Neural Networks. *arXiv*.

Pearlmutter, B. A. (1989). Learning State Space Trajectories in Recurrent Neural Networks. *Neural Computation*, 1(2):263–269.

Poggio, T. (2012). The Levels of Understanding framework, revised. *Computer Science and Artifi cialIntelligence Laboratory Technical Report*.

Quiroga, R. Panzeri, S. (2009). Extracting information from neuronal populations: information theory and decoding approaches. *Nature reviews. Neuroscience*, 10(3):173–85.

Quiroga, R. Q., Reddy, L., Koch, C., Fried, I. (2007). Decoding Visual Inputs From Multiple Neurons in the Human Temporal Lobe. *Journal of Neurophysiology*, 98(4):1997–2007.

Ramón y Cajal, S. (1894). The Croonian Lecture: La Fine Structure des Centres Nerveux. *Proceedings of the Royal Society of London*, 55:444–468.

Rao, R. P. N. Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87.

Robnik-Sikonja, M. Kononenko, I. (2008). Explaining Classifications For Individual Instances. *IEEE Transactions on Knowledge and Data Engineering*, 20(5):589–600.

Rochester, N., Holland, J., Haibt, L., Duda, W. (1956). Tests on a cell assembly theory of the action of the brain, using a large digital computer. *IEEE Transactions on Information Theory*, 2(3):80–93.

Rojas, R. (1996). Neural networks: a systematic introduction. *Neural Networks*, page 502.

Rosenblatt, F. (1957). The Perceptron A Perceiving and Recognizing Automaton. *Report 85-460-1, Cornell Aeronautical Laboratory*.

Rumelhart, D. E., Hinton, G. E., Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

Saxe, A. M., McClelland, J. L., Ganguli, S. (2013). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *CoRR*, abs/1312.6:1–22.

Schäfer, A. M. Zimmermann, H.-G. (2007). Recurrent Neural Networks are universal approximators. *International journal of neural systems*, 17(4):253–263.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117.

Schneidman, E., Bialek, W., Berry, M. J. (2003). Synergy, Redundancy, and Independence in Population Codes. *Journal of Neuroscience*, 23(37).

Scholte, H. S., Losch, M. M., Ramakrishnan, K., de Haan, E. H. F., Bohte, S. M. (2017). Visual pathways from the perspective of cost functions and multi-task deep neural networks. *arXiv*.

Schreiber, T. (2000). Measuring Information Transfer. *Physical Review Letters*, 85(2):461–464.

Schultz, W. (2016). Dopamine reward prediction error coding. *Dialogues in clinical neuroscience*, 18(1):23–32.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27:379–423.

Spoerer, C., McClure, P., Kriegeskorte, N. (2017). Recurrent Convolutional Neural Networks: A Better Model Of Biological Object Recognition Under Occlusion. *bioRxiv*.

Spratling, M. W. (2015). A review of predictive coding algorithms. *Brain and Cognition*, pages 1–8.

Stanley, K. O. Miikkulainen, R. (2002). Efficient reinforcement learning through evolving neural network topologies. *Proceedings of the Genetic and Evolutionary Computation Conference*.

Sussillo, D. (2014). Neural circuits as computational dynamical systems. *Current Opinion in Neurobiology*, 25:156–163.

Sussillo, D. Barak, O. (2013). Opening the Black Box: Low-Dimensional Dynamics in High-Dimensional Recurrent Neural Networks. *Neural Computation*, 25(3):626–649.

Szegedy, C., Zaremba, W., Sutskever, I. (2013). Intriguing properties of neural networks. *arXiv*, pages 1–10.

Teuber, H. L. Hans-Lukas (1955). Physiological Psychology. *Annual Review of Psychology*, 6(1):267–296.

Thornton, C. (2017). Predictive processing simplified: The infotropic machine. *Brain and Cognition*, 112:13–24.

Thorpe, S., Fize, D., Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, 381(6582):520–522.

Tiesinga, P., Fellous, J.-M., Sejnowski, T. J. (2008). Regulation of spike timing in visual cortical circuits. *Nature reviews. Neuroscience*, 9(2):97–107.

Turing, A. M. (1937). On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, s2-42(1):230–265.

van Gerven, M. A. (2016). A primer on encoding models in sensory neuroscience. *Journal of Mathematical Psychology*.

Vapnik, V. Lerner, A. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24:774–780.

Vasu, M. C. Izquierdo, E. J. (2017). Evolution and Analysis of Embodied Spiking Neural Networks Reveals Task-Specific Clusters of Effective Networks. *arXiv*.

Wadhwa, A. Madhow, U. (2016). Learning Sparse, Distributed Representations using the Hebbian Principle. *arXiv*.

Walsh, V. Cowey, A. (2000). Transcranial magnetic stimulation and cognitive neuroscience. *Nature Reviews Neuroscience*, 1(1):73–80.

Watson, J. Crick, F. (1953). Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*.

Werbos, P. J. (1974). Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences — BibSonomy. *Harvard University Ph.D. Thesis*.

Werbos, P. J. (1982). Applications of advances in nonlinear sensitivity analysis. In *System Modeling and Optimization*, pages 762–770. Springer-Verlag, Berlin/Heidelberg.

Yamins, D. DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 111(23):8619–24.

Yuste, R. (2015a). From the neuron doctrine to neural networks. *Nature Reviews Neuroscience*, 16.

Yuste, R. (2015b). On testing neural network models. *Nature Reviews Neuroscience*, 0594:24503.

Zeiler, M. D. Fergus, R. (2013). Visualizing and Understanding Convolutional Networks. *arXiv*.

Zenke, F. Ganguli, S. (2017). SuperSpike: Supervised learning in multi-layer spiking neural networks. *arXiv*.

Zintgraf, L., Cohen, T., Adel, T., Welling, M. (2016). Visualizing Deep Neural Network Decisions: Prediction Difference Analysis. *arXiv*, pages 1–12.