Deep Neural Networks in Computational Neuroscience

Tim C Kietzmann[1], Patrick McClure[1], & Nikolaus Kriegeskorte[1,2]

[1] MRC Cognition and Brain Science Unit, Cambridge, UK
[2] Department of Psychology, Columbia University

Keywords: deep neural networks, deep learning, convolutional neural networks, objective functions, recurrence, black box, levels of abstraction, modelling the brain, input statistics, biological detail

Summary

The goal of computational neuroscience is to find mechanistic explanations of how the nervous system processes information to support cognitive function and behaviour. At the heart of the field are its models, i.e. mathematical and computational descriptions of the system being studied. These models typically map sensory stimuli to neural responses and/or neural to behavioural responses and range for simple to complex. Recently, deep neural networks (DNNs), using either feedforward and recurrent architectures, have come to dominate several domains of artificial intelligence (AI). As the term "neural network" suggests, these models are inspired by biological brains. However, current DNN models abstract from many details of biological neural networks. Their abstractions contribute to their computational efficiency, enabling to perform complex feats of intelligence, ranging from perceptual tasks (e.g. visual object and auditory speech recognition) to cognitive tasks (e.g. machine translation), and on to motor control tasks (e.g. playing computer games or controlling a robot arm). In addition to their ability to model complex intelligent behaviours, DNNs have been shown to predict neural responses to novel sensory stimuli that cannot be predicted with any other currently available type of model. DNNs can have millions of parameters (connection strengths), which are required to capture the domain knowledge needed for task performance. These parameters are often set by task training using stochastic gradient descent. The computational properties of the units are the result of four directly manipulable elements: input statistics, network structure, functional objective, and learning algorithm. The advances with neural nets in engineering provide the technological basis for building task-performing models of varying degrees of biological realism that promise substantial insights for computational neuroscience.

The brain is a deep neural network

The goal of computational neuroscience is to find mechanistic explanations for how the nervous system processes information to support cognitive function and adaptive behaviour. Computational models, i.e. mathematical and computational descriptions of component systems, capture the mapping of sensory input to neural responses and explain representational transformations, neuronal dynamics, and the way the brain controls behaviour.

The overarching challenge is therefore to define models that explain neural measurements as well as complex adaptive behaviour. Computational neuroscientists have had early successes with shallow, linear-nonlinear "tuning", modelling lower-level sensory processing. Relatively shallow models have fuelled progress in the past and will continue to do so. Yet, the brain is a deep recurrent neural network that exploits multistage non-linear transformations and complex dynamics. It therefore seems inevitable that computational neuroscience will come to rely increasingly on deep recurrent models. The need for multiple stages of nonlinear computation has long been appreciated in the domain of vision, by both experimentalists (Hubel & Wiesel, 1959, 1962) and theorists (Fukushima, 1980; Lecun & Bengio, 1995; Riesenhuber & Poggio, 1999; G. Wallis & Rolls, 1997).

The traditional focus on shallow models was motivated both by the desire for simple explanations and by the difficulty of fitting complex models. Hand-crafted features, which laid the basis of modern computational neuroscience (Jones & Palmer, 1987), do not carry us beyond restricted lower-level tuning functions. As an alternative approach, researchers started directly using neural data to fit model parameters (Dumoulin & Wandell, 2008; M. C.-K. Wu, David, & Gallant, 2006). Despite its elegance, importance, and success, this approach is limited by the amount of neural observations that can be collected from a given system. Even with neural measurement technology advancing rapidly (multi-site array recordings, two-photon imaging, or neuropixels, to name just a few), the amount of recordable data does not provide enough constraints to fit realistically complex neural models. For instance, while novel measurement techniques may record separately from hundreds of individual neurons, and the number of stimuli used may approach 10,000, the numbers of parameters in deep neural networks (DNNs) are many orders of magnitude larger (for instance, the influential object recognition network "AlexNet" has 60 million parameters (Krizhevsky, Sutskever, & Hinton, 2012), a more recent object recognition network, VGG-16, has 138 million parameters (Simonyan & Zisserman, 2015)).

An important lesson in the history of AI is that intelligence requires a lot of domain knowledge. Transferring this knowledge into the model parameters through the bottleneck of neural measurements alone is too inefficient for complex models. A key insight that opened the path for the use of very complex models for prediction of neural responses is the idea that rather than fitting parameters based on neural observations, DNNs could instead be trained to perform relevant behaviour in the real world. This approach brings machine learning to bear on models for computational neuroscience, enabling researchers to constrain the model parameters via task training. In the domain of vision, for instance, category-labelled sets of training images can easily be assembled using web-based technologies, and the amount of available data can therefore be

expanded more easily than for measurements of neural activity. Of course a model trained to excel at a relevant task (such as object recognition, if we are trying to understand the computations in the primate ventral stream) might not be able to explain neural data. Testing which model architectures, input statistics, and learning objectives yield the best predictions of neural activity in novel experimental conditions (e.g. a set of images that has not been used in fitting the parameters) is a powerful way to learn about the computational mechanisms that might underlie the neural responses. The combined use of task training- and neural data enables us to build complex models with massive knowledge about the world in order to explain how biological brains implement cognitive function. Deep learning provides a very efficient tool to transfer this knowledge into the parameters of the model.

Brain-inspired neural network models are revolutionising artificial intelligence and exhibit rich potential for computational neuroscience

Neural network models inspired by biological brains have become a central class of models in machine learning (Figure 1). Driven by optimizing task-performance, they developed and improved model architectures, hardware and training schemes that eventually led to today's high-performance deep neural network models. These models have revolutionised several domains of AI, including computer vision (LeCun, Bengio, & Hinton, 2015). Starting with the seminal work by Krizhevsky et al (2012) , who won the ImageNet competition in visual object recognition by a large margin, deep neural networks now dominate computer vision (He, Zhang, Ren, & Sun, 2016; Simonyan & Zisserman, 2015; Szegedy et al., 2015), and drove reinforcement learning (Lange & Riedmiller, 2010; Mnih et al., 2015), speech-recognition (Sak, Senior, & Beaufays, 2014), machine translation (Sutskever, Vinyals, & Le, 2014; Y. Wu et al., 2016), and many other domains to unprecedented performance levels. In terms of visual processing, deep convolutional, feed-forward networks now achieve human-level classification performance (VanRullen, 2017).

Although inspired by biology, current DNNs abstract from all but the most essential features of biological neural networks. They are composed of simple units that typically compute a linear combination of their inputs and pass the result through a static nonlinearity (e.g. setting negative values to zero). To what extent they can nevertheless bring insights to computational neuroscience is controversial (Kay & Weiner, 2017; Kriegeskorte, 2015; VanRullen, 2017). Optimized to perform, DNNs differ substantially from biological neural networks, but they also exhibit architectural similarities. Consider the particularly successful variant of feedforward convolutional neural networks. Inspired by biological vision, these networks process images through a sequence of visuotopic representations. Each unit "sees" a restricted local region of the visuotopic map in the previous layer (its receptive field). Moreover, units are grouped into sets that detect the same visual feature all over the image (feature maps). The units within a feature map jointly learn a single connection weight template. The restriction to local receptive fields and sharing of weights among units in the same feature map greatly reduce the number of parameters that need to be learned. Like the primate visual system, convolutional neural

networks perform a deep cascade of non-linear computations, their neurons exhibit spatially restricted receptive fields that increase in size, invariance, and complexity along the hierarchical levels and similar feature detectors exist for different spatial locations in a given layer (although this is only approximately true in the primate brain). At the same time, however, these models are simplified in radical ways. They do typically not include lateral or top-down connections, compute continuous outputs (real numbers that could be interpreted as firing rates) rather than spikes. The list of features of biological neural networks not captured by these models is endless.

Despite abstracting from many features of biology, deep convolutional neural networks predict functional signatures of primate visual processing at multiple hierarchical levels. Trained to recognise objects, they develop V1-like receptive fields in early layers, and are predictive of single cell recordings in macaque IT (Cadieu et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014a; for reviews see Kriegeskorte, 2015; Yamins et al., 2014; Yamins & DiCarlo, 2016). In particular, the explanatory power of DNNs was on par with the performance of linear prediction based on an independent set of IT neurons and beyond linear predictions based directly on the category labels on which the networks were trained (Yamins et al., 2014). DNNs thereby constitute the only model class in computational neuroscience that is capable of predicting responses to novel images in IT with reasonable accuracy. DNNs explain about 50% of the variance of windowed spike counts in IT across individual images (Yamins et al., 2014), a performance level comparable to that achieved with Gabor models in V1 (Olshausen & Field, 2005). DNN modelling has also been shown to improve predictions of intermediate representations in area V4 over alternative models (Yamins & DiCarlo, 2016). This indicates that, in order to solve the task, the trained network transforms the image through a similar sequence of intermediate representations as the primate brain.

In human neuroscience similarly, DNNs proved capable of predicting representations measured with functional magnetic resonance imaging across multiple levels of processing in a hierarchical fashion: lower network levels better predict lower level visual representations, and subsequent, higher-levels better predict activity in higher- more anterior cortical areas (Güçlü & van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014a). In line with results from macaque IT, DNNs were furthermore able to explain within-category neural similarities, despite being trained on a categorization task (Khaligh-Razavi & Kriegeskorte, 2014a). At a lower spatial, but higher temporal resolution, DNNs have also been shown to be predictive of visually evoked magnetoencephalography (MEG) data (Cichy, Khosla, Pantazis, & Oliva, 2016; Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016; Fritsche, G, Schoffelen, Bosch, & Gerven, 2017). On the behavioural level, deep networks exhibit similar behaviour (Hong, Yamins, Majaj, & DiCarlo, 2016; Kheradpisheh, Ghodrati, Ganjtabesh, & Masquelier, 2016a, 2016b; Kubilius, Bracci, & Op de Beeck, 2016; Majaj, Hong, Solomon, & DiCarlo, 2015) and are currently the best-performing model in explaining human eye-movements in free viewing paradigms (Kümmerer, Theis, & Bethge, 2014). These early examples clearly illustrate the power of DNN models for computational neuroscience.

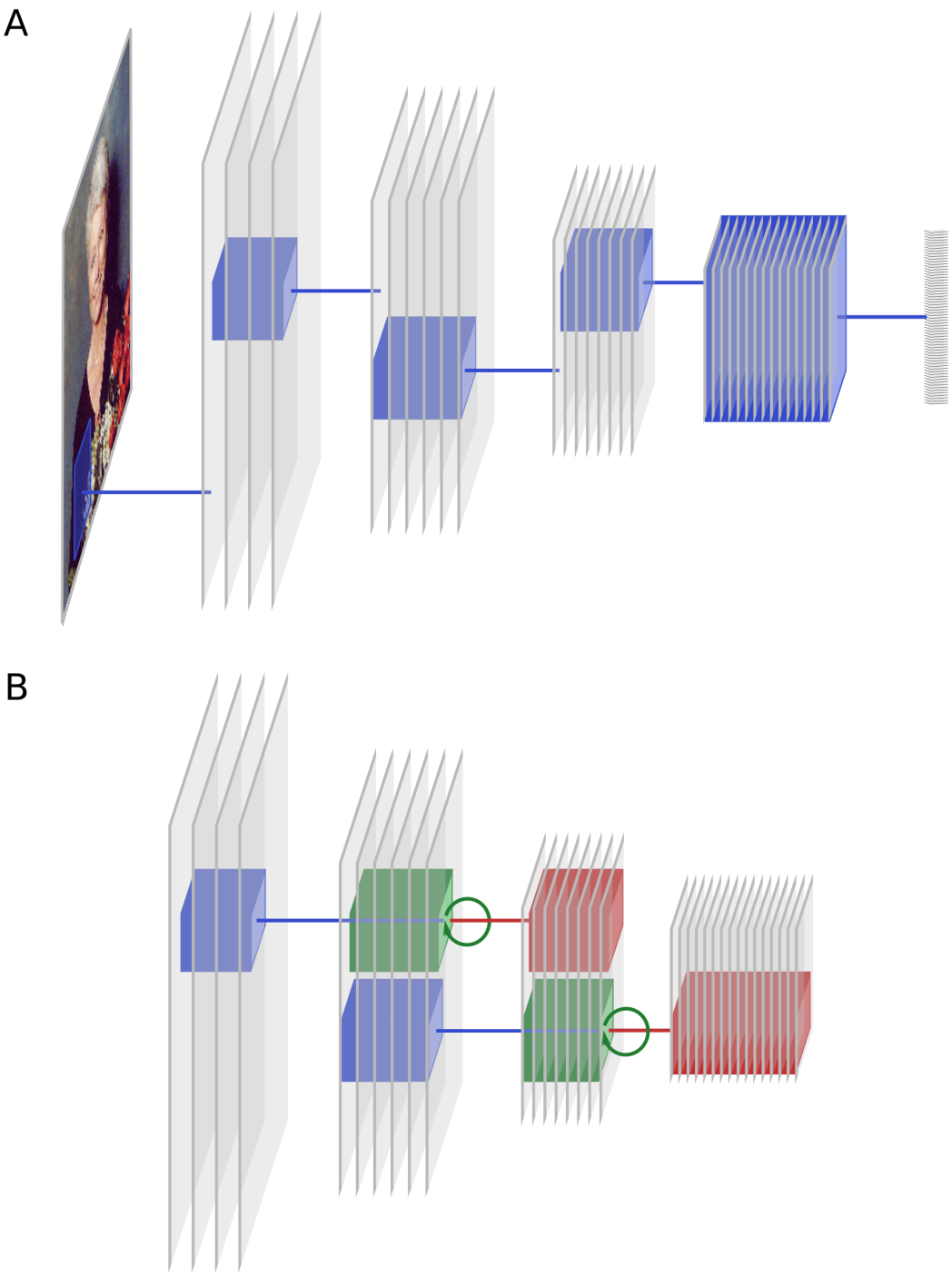

Figure 1. Convolutional neural network structure. (A) An example feed forward convolutional neural network (CNN) with 3 convolutional layers followed by a fully-connected layer. Bottom-up filters for selected neurons are illustrated with blue boxes. (B) The bottom-up (blue), lateral (green), and top-down (red) filters for two example neurons in different layers of a recurrent convolutional neural network (RCNN).

Deep neural network models can be tested with brain and behavioural data

DNNs are typically trained to optimize external behavioural objectives rather than being derived from neural data. Thus, model testing with activity measurements is crucial to assess how well a network matches cortical responses. DNNs excel at task performance, but even human-level performance does not imply that the underlying computations employ the same mechanisms. Fortunately, computational neuroscience has a rich toolbox at its disposal that allows researchers to probe even highly complex models, such as DNNs (Diedrichsen & Kriegeskorte, 2017).

One such tool are encoding models, which use external, fixed feature spaces in order to model neural responses across a large variety of experimental conditions (e.g. different stimuli, Figure 2A). The underlying idea is that if the model and the brain compute the same features, then linear combinations of the model features should enable successful prediction of the neural responses for independent experimental data (Naselaris, Kay, Nishimoto, & Gallant, 2011). For visual representations, the model feature space can derive from simple filters, such as Gabor-wavelets (Kay, Naselaris, Prenger, & Gallant, 2008), from human labelling of the stimuli (Huth, Nishimoto, Vu, & Gallant, 2012; Mitchell et al., 2008; Naselaris, Prenger, Kay, Oliver, & Gallant, 2009), or responses in different layers of a DNN (Agrawal, Stansbury, Malik, & Gallant, 2014; Güçlü & van Gerven, 2015).

Probing the system on the level of multivariate response patterns, representational similarity analysis (RSA, Kriegeskorte & Kievit, 2013; Kriegeskorte, Mur, & Bandettini, 2008; Nili et al., 2014) provides another approach to comparing internal representations in DNNs and the brain (Figure 2B). RSA characterizes the representational geometry in a given system by the representational pattern dissimilarities among the stimuli. A model representation is considered similar to a brain representation to the degree that it emphasizes the same distinctions among the stimuli. Stimulus-by-stimulus representational dissimilarity matrices can be directly compared between brain regions and model layers, side-stepping the problem of defining the correspondency mapping between the units of the model and the channels of brain-activity measurement (e.g. voxels in fMRI, (Carlin, Calder, Kriegeskorte, Nili, & Rowe, 2011; Guntupalli, Wheeler, & Gobbini, 2016; Khaligh-Razavi & Kriegeskorte, 2014a; Kietzmann, Swisher, König, & Tong, 2012), single-cell recordings (Kriegeskorte et al., 2008; Leibo, Liao, Freiwald, Anselmi, & Poggio, 2017; Tsao, Moeller, & Freiwald, 2008), M/EEG data (Cichy, Pantazis, & Oliva, 2014; Kietzmann, Gert, Tong, & König, 2017), and behavioural measurements including perceptual judgments (Mur et al., 2013).

On the behavioural level, recognition performance (Cadieu et al., 2014; Hong et al., 2016; Majaj et al., 2015), perceptual confusions, and illusions provide valuable clues as to how representations in brains and DNNs may differ. For instance, it can be highly informative to understand the detailed patterns of errors (Walther, Caddigan, Fei-Fei, & Beck, 2009) and reaction times across stimuli, which may reveal subtle functional differences between systems that exhibit the same overall level of task performance. Visual metamers (Freeman & Simoncelli, 2011; T. S. A. Wallis, Bethge, & Wichmann, 2016) provide a powerful tool to test for similarities in internal representations across systems. Given an original image, a modified version is created that matches the original in the model representation (for instance, a layer of a DNN), while features that do not

change the representation are altered. If the human brain processed the stimuli through the same stages, it should similarly be insensitive to the two stimuli that are indistinguishable ("metameric") to the model. Conversely, an adversarial example (Goodfellow, Shlens, & Szegedy, 2015; Nguyen, Yosinski, & Clune, 2015) is a minimal modification of an image that elicits a different category label from a DNN. For convolutional feedforward networks, minimal changes to an image (say of a bus), which are imperceptible to humans, suffice lead the model to classify the image incorrectly (say as an ostrich). Adversarial examples can be generated using the backpropagation algorithm down to the level of the image, to find the gradients in image space that change the classification output. This method requires omniscient access to the system, making it impossible to perform a fair comparison with biological brains, which might likewise be confused by stimuli designed to exploit the idiosyncratic aspects (Kriegeskorte, 2015). The more general lesson for computational neuroscience is that metamers and adversarial examples provide methods for designing stimuli for which different representations disagree maximally. This may enable us in the future to optimise our power to adjudicate between alternative models experimentally.

Ranging across levels of description and modalities of brain-activity measurement, from responses in single neurons, to array recordings, fMRI and MEG data, and behavioural responses, the above methods enable computational neuroscientists to investigate the similarities and differences between brains and DNNs. Future studies can explore a wide range of model units and network architectures, adding features consistent with neurobiology so as to best predict brain activity and behaviour.

Drawing theoretical insight from complex models

Deep learning has transformed machine learning and only recently found its way back into computational neuroscience, where it originated. Despite their high performance, DNNs have met with scepticism regarding their explanatory value as models of brain information processing (e.g. Kay & Weiner, 2017). One of the arguments commonly put forward is that DNNs merely exchange one impenetrably complex system with another (the "black box" argument). That is, while DNNs may be able to predict neural data, researchers now face the problem of understanding what exactly the network is doing.

The black box argument is best appreciated in historical context. Shallow models are easier to understand and supported by stronger mathematical results. For example, the weight template of a linear-nonlinear model can be directly visualised and understood in relation to the concept of an optimal linear filter. Simple models can furthermore enable researchers to understand the role of each individual parameter. Overall, a model with fewer parameters is considered more parsimonious as a theoretical account.

It is certainly true that simpler models should be preferred over models with excessive degrees of freedom. Many seminal explanations in neuroscience have been derived from simple models. This argument only applies, however, if the two models provide similar predictive power. Models should be as simple as possible, but no simpler. Because the brain is a complex system with billions of parameters (presumably containing the domain knowledge required for adaptive behaviour) and complex

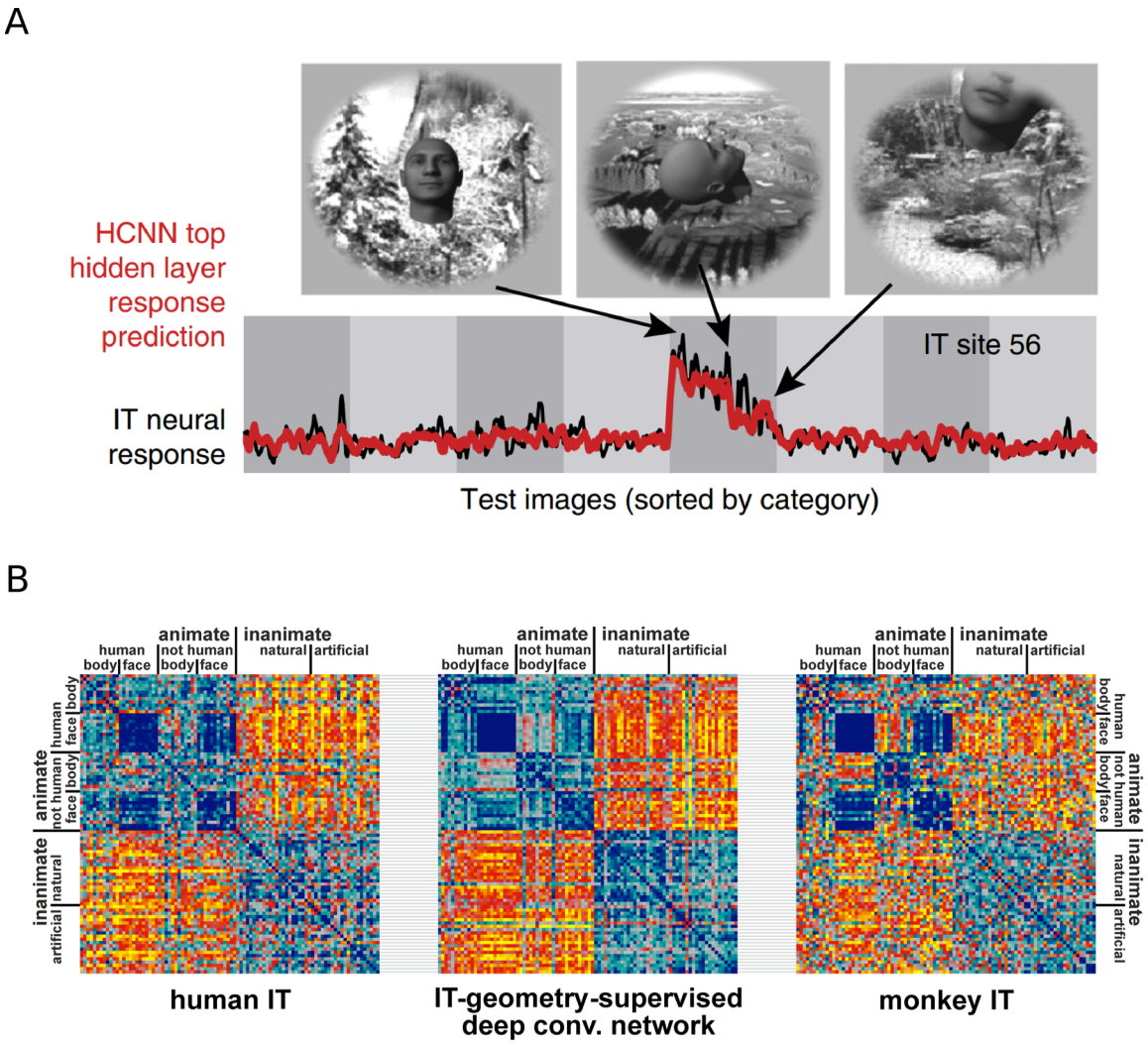

Figure 2. Testing the internal representations of DNNs against neural data. (A) An example of neuron-level encoding with a convolutional neural network (adapted from Yamins & DiCarlo, 2016). (B) Comparing the representational geometry of a trained CNN to human and monkey brain activation patterns using representation-level similarity analysis (adapted from Khaligh-Razavi & Kriegeskorte, 2014b).

dynamics (which implement perceptual inference, cognition, and motor control), computational neuroscience will eventually need complex models. The field has to find ways to draw insight from such models. One way to draw insight from complex models is to consider their constraints at a higher level of abstraction. The computational properties of DNNs are the result of four manipulable elements: the network architecture, the *input statistics*, the *functional objective*, and the *learning algorithm*.

A worthwhile thought experiment for neuroscientists is to consider what cortical representations would develop if the world were different. Governed by different input statistics, for instance, a different distribution of category occurrences or different temporal dependency structure, the brain may develop quite differently. This knowledge

would provide us with principal insights into the objectives that it tries to solve during development. Deep learning allows computational neuroscientists to make this thought experiment a simulated reality. Investigating which aspects of the simulated world are crucial to render the learned representations more similar to the brain thereby serves an essential function in understanding of representational characteristics.

In addition to experiments with different input statistics, the network architecture can be altered to test how anatomical structure gives rise to computational function, and which features of the biological brain are required to explain a given neural phenomenon. For instance, it can be asked whether neural responses in a given paradigm are best explained by a feed-forward or a recurrent network architecture. Moreover, starting from an abstract level, biological details can be integrated into DNNs in order to see which ones prove to be required ingredients for predicting neural responses and behaviour. Current DNNs derive their power from bold abstractions. Although complex in terms of their parameter count, they are simple in terms of their component mechanisms. Biological brains draw from a richer set of dynamical primitives. It will be interesting to see to what extent incorporating more biologically inspired mechanisms can further enhance the power of DNNs and their ability to explain neural activity and animal behaviour.

Given input statistics and architecture, the missing determinants that transform the randomly initialised model into a trained DNN are the objective function and the learning algorithm. The idea of normative approaches is that neural representations in the brain can be understood as being optimized with regard to one or many overall objectives. These define what the brain should compute, in order to provide the basis for successful behaviour. While experimentally difficult to investigate, deep learning based on different cost functions allows researchers to ask the directly related inverse question: what cost functions need to be optimized such that the resulting internal representations best predict neural data? Various objectives have been suggested in both the neuroscience and machine learning community. Feed-forward convolutional DNNs are often trained with the objective to minimize classification error (Krizhevsky et al., 2012; Simonyan & Zisserman, 2015; Yamins & DiCarlo, 2016). This focus on classification performance has proven quite successful, leading researchers to observe an intriguing correlation: classification performance is positively related to the ability to predict neural data (Khaligh-Razavi & Kriegeskorte, 2014a; Yamins et al., 2014). That is, the better the network performed on a given image set, the better it could predict neural data, although the latter was not part of the training objective.

The objective to minimize classification error in a DNN for visual object recognition requires millions of labelled training images. Although the finished product, the trained DNN provides the best current model of ventral stream visual processing, the process by which the model is obtained is not biologically plausible. The image labels are best viewed as a crutch for semantics, which replaces the contribution of the rest of the brain and the body in interaction with a complex dynamic environment. Objective functions from the unsupervised domain have been suggested, which would allow the brain (and DNNs) to create error signals without external feedback. One influential suggestion is that neurons in the brain aim at an efficient sparse code, while faithfully representing the external information (Olshausen & Field, 1996; Simoncelli & Olshausen, 2001). Similarly, compression-based objectives aim to represent the input with as few

neural dimensions as possible. Autoencoders are one example of this coding principle (Hinton & Salakhutdinov, 2006).

Harnessing information from the temporal domain, the temporal stability or slowness objective is based on the insight that latent variables that vary slowly over time are useful for adaptive behaviour. Neurons should therefore detect the underlying, slowly changing signals, while disregarding fast changes likely due to noise, potentially simplifying readout from downstream neurons (Berkes & Wiskott, 2005; Földiák, 1991; C. Kayser, Körding, & König, 2003; Christoph Kayser, Einhäuser, Dümmer, König, & Körding, 2001; Körding, Kayser, Einhäuser, & König, 2004; Rolls, 2012; Wiskott & Sejnowski, 2002). Stability can be optimized across layers in hierarchical systems, if each subsequent layer tries to find an optimally stable solution from the activation profiles in previous layer. This approach was shown to lead to invariant codes for object identity (Franzius, Wilbert, & Wiskott, 2008) and viewpoint-invariant place-selectivity (Franzius, Sprekeler, & Wiskott, 2007; Wyss, König, & Verschure, 2006). Experimental evidence in favour of the temporal stability objective has been provided by electrophysiological and behavioural studies (N. Li & DiCarlo, 2008, 2010; G. Wallis & Bülthoff, 2001).

Many implementations of classification, sparseness and stability objectives ignore the action repertoire of the agent. Yet, different cognitive systems living in the same world may exhibit different neural representations because the requirements to optimally support action may differ. Deep networks optimizing the predictability of the sensory consequence (Weiller, Märtin, Dähne, Engel, & König, 2010), or cost of a given action (Mnih et al., 2015) have started incorporating the corresponding information. Finally, there does not have to be one true objective that the brain optimizes, as neural cost functions are not necessarily constant across regions or time (Marblestone, Wayne, & Kording, 2016).

As a result, one way to draw theoretical insights from DNN models is to explore what architectures, input statistics, objective functions, and learning algorithms yield models predictive of neural activity and behaviour. This approach does not elucidate the role of individual units or connections, but it can reveal what features of biological structure support what aspects of a system's function and what objectives the biological system might be optimised for.

In addition to contextualising the black box in this way, we can also open the black box and look inside. Given a model that accounts for neural activity and behaviour, much is won. Unlike a biological brain, a model is entirely accessible to scrutiny and manipulation, enabling, for example, high-throughput "in silico" electrophysiology. One method for visualizing a unit's preferences is to approximately undo the operations performed by a convolutional DNN (Zeiler & Fergus, 2014) to visualise what image features drive a given unit deep in a neural network in the context of a particular image. This results in visualisations such as those shown in Figure 3. A closely related technique is to use backpropagation (Rumelhart, Hinton, & Williams, 1986) to calculate the change in the input needed to drive or inhibit the activation of any unit in a DNN (Simonyan & Zisserman, 2015; Yosinski, Clune, Nguyen, Fuchs, & Lipson, 2015). We can select an image that strongly drives the unit and compute the gradient in image space that corresponds to enhancing the unit's activity even further. The gradient image shows how small adjustments to the pixels affect the activity of the unit. For example, if the image strongly driving the unit is a person next to a car, corresponding the gradient image might

reveal that it is really the face of the person driving the unit's response. In that case, the gradient image would deviate from zero only in the region of the face and adding the gradient image to the original image would accentuate the facial features. To understand the unit's response, we might have to look at its gradient in image space for many different test images to get a sense of the orientation of its tuning surface around multiple reference points (test images).

Backpropagation can also be used to iteratively optimise images to strongly drive a particular unit, starting from a noise image. This yields complex psychedelic looking patterns containing features and forms, the network has learned through its task training. It is important to note that the tuning function of a unit deep in a network cannot be characterised by a single visual template. If it could, there would be no need for multiple stages of nonlinear transformation. However, visualizations of receptive field properties provide intuitions about the neuronal selectivity at different layers or time-points.

In summary, in silico electrophysiology enables researchers to measure and manipulate every single neuron, if required. In addition, researchers can gain an understanding at a more abstract level, by observing the effects of predictive performance of changes to the architecture, input statistics, objective function, and learning algorithm.

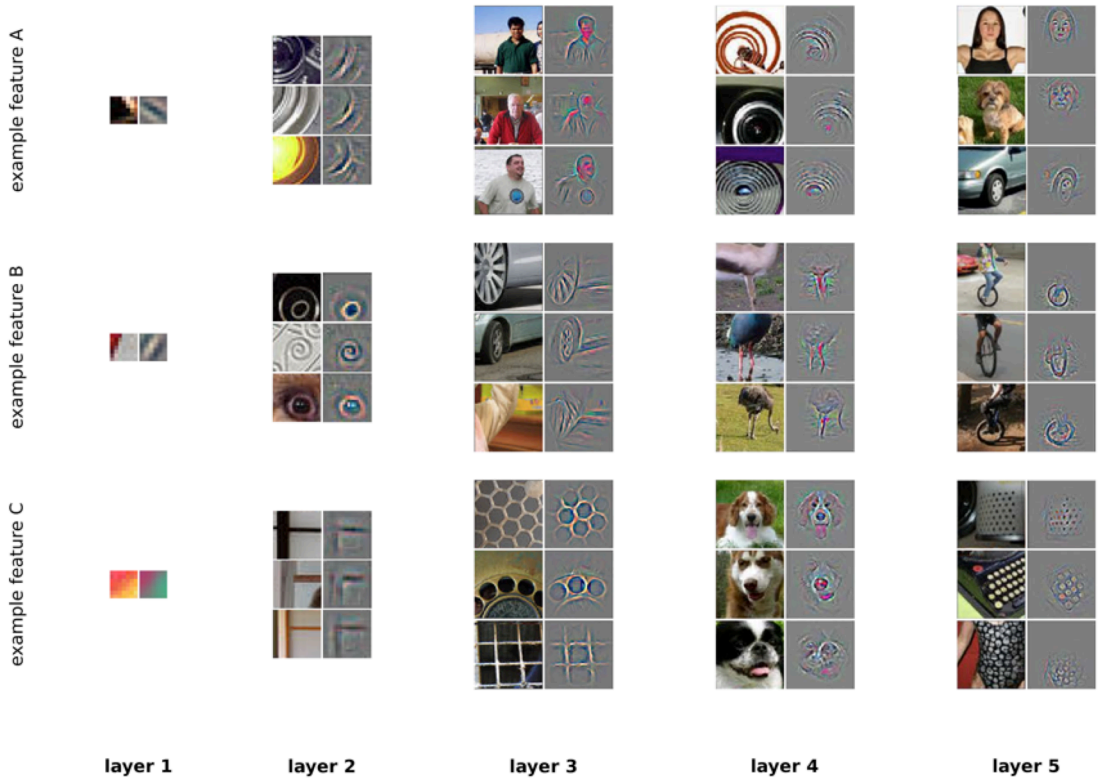

Figure 3. Visualizing the preferred features of internal neurons. Activations in a random subset of feature maps across layers for strongly driving ImageNet images projected down to pixel space (adapted from Zeiler & Fergus, 2014).

What neurobiological details matter to brain computation?

A second concern about DNNs is that they abstract too much from biological reality to be of use as models for neuroscience. Whereas the black box argument states that DNNs are too complex, the biological realism argument states that they are too simple. Both arguments have merit. It is conceivable that a model is simultaneously to simple (in some ways) and too complex (in other ways). However, this raises a fundamental question: Which features of the biological structure should be modelled and which omitted to explain brain function?

Abstraction is the essence modelling and is the driving force of understanding. If the goal of computational neuroscience is to understand brain *computation*, then we should seek the simplest models that can explain task performance and predict neural data. The elements of the model should map onto the brain at some level of description. However, what biological elements must be modelled is an empirical question. Large-scale models should enable an exploration of the level of detail required (Eliasmith & Trujillo, 2014). DNNs are important not because they capture the biological features that matter to brain computation, but because they provide a minimal functioning starting point for exploring what biological details matter to brain computation. If, for instance, spiking models outperformed rate-coding models at explaining neural activity and task performance (for example in tasks requiring probabilistic inference (Buesing, Bill, Nessler, & Maass, 2011), then this would be strong evidence in favour of spiking models. Convolutional DNNs like AlexNet (Krizhevsky et al., 2012), and VGG (Simonyan & Zisserman, 2015) were built to optimise performance, rather than biological plausibility. However, these models draw from a history of neuroscientific insight and share many qualitative features with the primate ventral stream. The defining property of convolutional DNN is the use of convolutional layers. These have two main characteristics: (1) local connections that define receptive fields and (2) parameter sharing between neurons across the visual field. Whereas spatially restricted receptive fields are a prevalent biological phenomenon, parameter sharing is biologically implausible. However, biological visual systems learn qualitatively similar sets of basis features in different parts of a retinotopic map, and similar results have been observed in models optimizing a sparseness objective (Güçlü & van Gerven, 2014; Olshausen & Field, 1996).

Moving toward greater biological plausibility with DNNs, locally connected layers that have receptive fields without parameter sharing were suggested (Uetz & Behnke, 2009). Researchers have already started exploring this type of DNN, which was shown to be very successful in face recognition (Sun, Wang, & Tang, 2015; Taigman, Ranzato, Aviv, & Park, 2014). One reason for this is that locally connected layers work best in cases where similar features are frequently present in the same visual arrangement, such as faces. In the brain, retinotopic organization principles have been proposed for higher-level visual areas (Levy, Hasson, Avidan, Hendler, & Malach, 2001), and similar organisation mechanisms may have led to faciotopy, the spatially stereotypical activation for facial features across the cortical surface in face-selective regions (Henriksson, Mur, & Kriegeskorte, 2015).

Another aspect in which convolutional AlexNet and VGG deviate from biology is the focus on feed-forward processing. Feedforward DNNs compute static functions, and

are therefore limited to modelling the feed-forward sweep of signal flow through a biological visual system. Yet, recurrent connections are a key computational feature in the brain, and represent a major research frontier in neuroscience. In the visual system, too, recurrence is a ubiquitous phenomenon. Recurrence is likely the source of representational transitions from global to local information (Matsumoto, Okada, Sugase-Miyamoto, Yamane, & Kawano, 2005; Sugase, Yamane, Ueno, & Kawano, 1999). The timing of signatures of facial identity (Barragan-Jason, Besson, Ceccaldi, & Barbeau, 2013; Freiwald & Tsao, 2010) and social cues, such as direct eye-contact (Kietzmann et al., 2017), point towards a reliance on recurrent computations. Finally, recurrent connections likely play a vital role in dealing with occlusion (Spoerer, McClure, & Kriegeskorte, 2017; Wyatte, Curran, & O'Reilly, 2012; Wyatte, Jilk, & O'Reilly, 2014) and object-based attention (Roelfsema, Lamme, & Spekreijse, 1998).

The first generation of DNNs focused on feed-forward, but the general class of DNNs can implement recurrence. By using lateral recurrent connections, DNNs can implement visual attention mechanisms (Z. Li, Yang, Liu, Wen, & Xu, 2017; Mnih, Heess, Graves, & Kavukcuoglu, 2014), and lateral recurrent connections can also be added to convolutional DNNs (Liang & Hu, 2015; Spoerer et al., 2017), which increases the effective receptive field size of each unit. In addition to local feedforward and lateral recurrent connections, the brain also uses local feedback, as well as long-range feedforward and feedback connections. While missing from the convolutional DNNs previously used to predict neural data, DNNs with these different connection types have been implemented (He et al., 2016; Liao & Poggio, 2016; Srivastava, Greff, & Schmidhuber, 2015). The field of recurrent convolutional DNNs is still in its infancy, and the effects of lateral and top-down connections on the representational dynamics in these networks, and their predictive power for neural data are yet to be fully explored. Nevertheless, recurrent connections are an exciting tool for computational neuroscience and will likely allow for insights into the recurrent computational dynamics of the brain.

Apart from architectural considerations, backpropagation, the most successful learning algorithm for DNNs, has classically been considered neurobiologically implausible. Rather than as a model of biological learning, backpropagation may be viewed as an efficient way to arrive at reasonable parameter estimates, which are then subject to further tests. That is, even if backpropagation is considered a mere technical solution, the resulting model may still be a good model of the dynamics in the system after learning. However, there is also a growing literature on biologically plausible forms of error-driven learning. If the brain does optimise cost functions during development and learning (which can be diverse, and supervised, unsupervised, or reinforcement-based), then it will have to use a form of optimization mechanism, such as stochastic gradient descent techniques. The current literature suggests several neurobiologically plausible ways in which the brain could adjust its internal parameters to optimise such objective functions (Lee, Zhang, Fischer, & Bengio, 2015; Lillicrap et al., 2016; O'Reilly, 1996; Whittington & Bogacz, 2015; Xie & Seung, 2003). These methods have been shown to allow deep spiking neural networks to learn simple vision tasks (Guergiuev, Lillicrap, & Richards, 2016). The brain might not be performing the exact algorithm of backpropagation, but it might have a mechanism for modifying synaptic weights in order to optimise one or many objective functions (Marblestone et al., 2016).

In addition to architectural considerations and optimization, there are other ways in which DNNs abstract from biological detail. For instance, DNNs are generally deterministic, while biological networks are stochastic. While much of this stochasticity is commonly thought to be noise, it has been hypothesized that this variability could code for uncertainty (Fiser, Berkes, Orbán, & Lengyel, 2010; Orban, Berkes, Fiser, & Lengyel, 2016). Furthermore, current recurrent convolutional DNNs often only run for a few time steps, and the roles of dynamical features found in biological networks, such as oscillations, are only beginning to be tested (Finger & König, 2013; Reichert & Serre, 2013; Siegel, Donner, & Engel, 2012). Another abstraction is the omission of spiking dynamics. However, DNNs with spiking neurons can be implemented (Tavanaei & Maida, 2016) and represent an exciting frontier of deep learning research. These considerations show that it would be hasty to judge the merits of DNNs based on the level of abstraction chosen in the first generation. The usage of DNNs in computational neuroscience is still in its infancy. Integration of biological detail will require close collaboration between modellers and experimental neuroscientists and anatomists.

Computational neuroscience comprises a wide range of models, defined at various *levels of biological and behavioural detail* (Figure 4). For instance, many conductance-based models contain large amounts of parameters to explain single or few neurons at great level of detail but are typically not geared towards behaviour. DNNs, at the other end of the spectrum, use their high number of parameters not to account for effects on the molecular level, but to achieve behavioural relevance, while accounting for overall neural selectivity. Explanatory merit is not only gained by biological realism (because this would render human brains the perfect explanation for themselves), nor does it directly follow from simplistic models that cannot account for complex animal behaviour. The space of models is continuous and neuroscientific insight works across multiple levels of explanation, following top-down and bottom-up approaches (Craver, 2009).

Conclusions

Deep neural networks have revolutionised machine learning and AI, and have recently moved back into computational neuroscience. These models reach human-level performance in certain tasks, and early experiments indicate that they are capable of capturing characteristics of cortical function that cannot be captured with shallow linear-nonlinear models. DNNs offer an intriguing new framework that enables computational neuroscientists to address fundamental questions about brain computation in the developing and adult brain.

DNNs will not replace shallow models, but rather enhance the investigative repertoire of the computational neuroscientist. Understanding neural computations is ultimately an interdisciplinary endeavour. Experimental neuroscientists will have to collaborate with machine learning researchers if we are to understand how the brain works. With computers approaching the brain in computational power, we are entering a truly exciting phase of computational neuroscience.

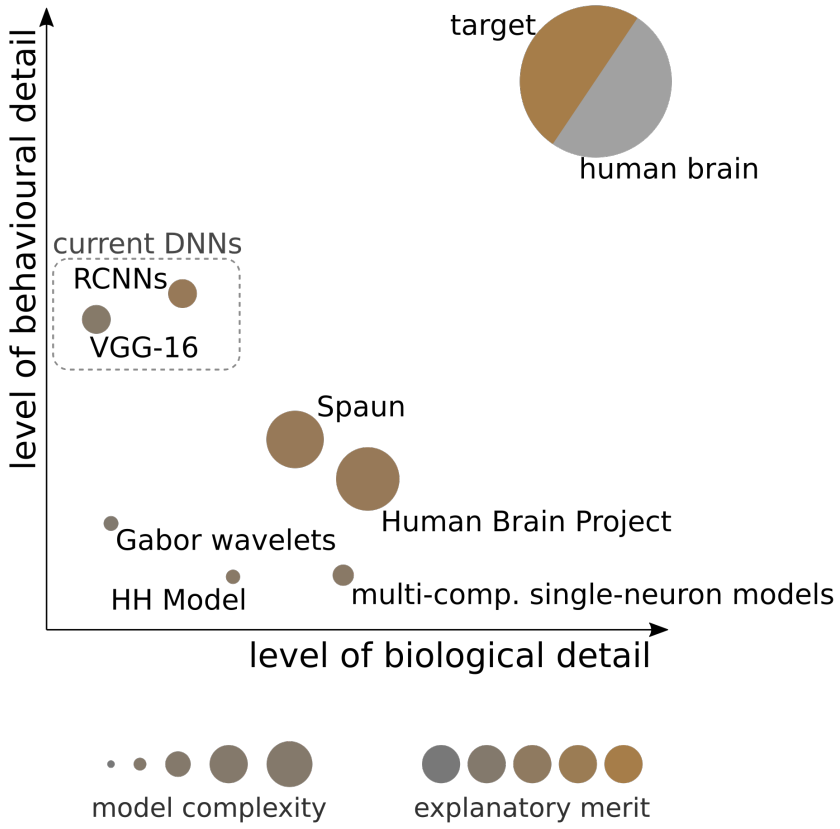

Figure 4. Cartoon overview of different models in computational neuroscience. Given computational constraints, models need to make simplifying assumptions. These can either be regarding the biological detail, or behavioral relevance of the model output. The explanatory merit of a model is not dependent on the exact replication of biological detail, but on its ability to provide insights into the inner workings of the brain at a given level of abstraction.

References

Agrawal, P., Stansbury, D., Malik, J., & Gallant, J. (2014). Pixels to Voxels: Modeling Visual Representation in the Human Brain. *arXiv Preprint arXiv: ...*, 1–15. Retrieved from http://arxiv.org/abs/1407.5104

Barragan-Jason, G., Besson, G., Ceccaldi, M., & Barbeau, E. J. (2013). Fast and Famous: Looking for the Fastest Speed at Which a Face Can be Recognized. *Frontiers in Psychology*, *4*(March), 100. http://doi.org/10.3389/fpsyg.2013.00100

Berkes, P., & Wiskott, L. (2005). Slow feature analysis yields a rich repertoire of complex cell properties. *Journal of Vision*, *5*, 579–602. http://doi.org/10.1167/5.6.9

Buesing, L., Bill, J., Nessler, B., & Maass, W. (2011). Neural dynamics as sampling: A model for stochastic computation in recurrent networks of spiking neurons. *PLoS Computational Biology*, *7*(11). http://doi.org/10.1371/journal.pcbi.1002211

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. a., … DiCarlo, J. J. (2014). Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *Arxiv*, *10*(12), 35. http://doi.org/10.1371/journal.pcbi.1003963

Carlin, J. D., Calder, A. J., Kriegeskorte, N., Nili, H., & Rowe, J. B. (2011). A Head View-Invariant Representation of Gaze Direction in Anterior Superior Temporal Sulcus. *Current Biology*, 1–5. http://doi.org/10.1016/j.cub.2011.09.025

Cichy, R. M., Khosla, A., Pantazis, D., & Oliva, A. (2016). Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *NeuroImage*, 1–13. http://doi.org/http://dx.doi.org/10.1101/032623

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Deep Neural Networks predict Hierarchical Spatio-temporal Cortical Dynamics of Human Visual Object Recognition. *arXiv*, 15. http://doi.org/10.1038/srep27755

Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience*, *17*, 455–462. http://doi.org/10.1038/nn.3635

Craver, C. (2009). Explaining the brain. Mechanisms and the mosaic unity of neuroscience 2007. *New York: Oxford University Press*. Retrieved from http://philpapers.org/rec/CRAETB-2

Diedrichsen, J., & Kriegeskorte, N. (2017). Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS Computational Biology*, 1–33. http://doi.org/10.1101/071472

Dumoulin, S. O., & Wandell, B. A. (2008). Population receptive field estimates in human visual cortex. *NeuroImage*, *39*(2), 647–660. http://doi.org/10.1016/j.neuroimage.2007.09.034

Eliasmith, C., & Trujillo, O. (2014). The use and abuse of large-scale brain models. *Current Opinion in Neurobiology*, *25*, 1–6. http://doi.org/10.1016/j.conb.2013.09.009

Finger, H., & König, P. (2013). Phase synchrony facilitates binding and segmentation of natural images in a coupled neural oscillator network. *Frontiers in Computational Neuroscience*, *7*(January), 195. http://doi.org/10.3389/fncom.2013.00195

Fiser, J., Berkes, P., Orbán, G., & Lengyel, M. (2010). Statistically optimal perception and learning: from behavior to neural representations. *Trends in Cognitive Sciences*.

http://doi.org/10.1016/j.tics.2010.01.003

Földiák, P. (1991). Learning Invariance from Transformation Sequences. *Neural Computation*, *3*, 194–200. http://doi.org/10.1162/neco.1991.3.2.194

Franzius, M., Sprekeler, H., & Wiskott, L. (2007). Slowness and sparseness lead to place, head-direction, and spatial-view cells. *PLoS Computational Biology*, *3*, 1605–1622. http://doi.org/10.1371/journal.pcbi.0030166

Franzius, M., Wilbert, N., & Wiskott, L. (2008). Invariant object recognition with slow feature analysis. In *Artificial Neural Networks-ICANN 2008* (pp. 961–970). Berlin Heidelberg: Springer. http://doi.org/10.1007/978-3-540-87536-9_98

Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Neuroscience*, *14*(9), 1195–201. http://doi.org/10.1038/nn.2889

Freiwald, W. A., & Tsao, D. Y. (2010). Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science*, *330*(6005), 845–51. http://doi.org/10.1126/science.1194908

Fritsche, M., G, U., Schoffelen, J., Bosch, S. E., & Gerven, M. A. J. Van. (2017). CNN-based Encoding and Decoding of Visual Object Recognition in Space and Time, (1980), 1–22. http://doi.org/10.1101/118091

Fukushima, K. (1980). Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. *Biological Cybernetics*, *46*, 193–202. Retrieved from http://link.springer.com/article/10.1007/BF00344251

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. *ICLR 2015*, 1–11. Retrieved from http://arxiv.org/abs/1412.6572

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. http://doi.org/10.1523/JNEUROSCI.5023-14.2015

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). http://doi.org/10.1371/journal.pcbi.1003724

Guergiuev, J., Lillicrap, T. P., & Richards, B. A. (2016). Deep learning with segregated dendrites. *arXiv Preprint*. http://doi.org/arXiv:1610.00161

Guntupalli, J., Wheeler, K., & Gobbini, M. (2016). Disentangling the Representation of Identity From Head View, 1–25.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, 770–778. http://doi.org/10.1007/s11042-017-4440-4

Henriksson, L., Mur, M., & Kriegeskorte, N. (2015). Faciotopy-A face-feature map with face-like topology in the human occipital face area. *Cortex*, *72*, 156–167. http://doi.org/10.1016/j.cortex.2015.06.030

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science*, *313*(5786), 504–507. http://doi.org/10.1126/science.1127647

Hong, H., Yamins, D. L., Majaj, N. J., & DiCarlo, J. J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature*

Neuroscience, *19*(4), 613–622. http://doi.org/10.1038/nn.4247

Hubel, D., & Wiesel, T. (1959). Receptive fields of single neurones in the cat's striate cortex. *Journal of Physiology*, *148*, 574–591. http://doi.org/10.1113/jphysiol.2009.174151

Hubel, D., & Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in cat's visual cortex. *The Journal of Physiology*, *160*, 106–154. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1359523/

Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, *76*(6), 1210–24. http://doi.org/10.1016/j.neuron.2012.10.014

Jones, J. P., & Palmer, L. a. (1987). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, *58*(6), 1233–1258. http://doi.org/citeulike-article-id:762473

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*(7185), 352–5. http://doi.org/10.1038/nature06713

Kay, K. N., & Weiner, K. S. (2017). Principles for models of neural information processing. *bioRxiv*, (April), 1–20.

Kayser, C., Einhäuser, W., Dümmer, O., König, P., & Körding, K. (2001). Extracting slow subspaces from natural videos leads to complex cells. *Artificial Neural Networks—ICANN*, 1075–1080. Retrieved from http://www.springerlink.com/index/MY6DNCHMAYQEYHRV.pdf

Kayser, C., Körding, K. P., & König, P. (2003). Learning the nonlinearity of neurons from natural visual stimuli. *Neural Computation*, *15*(8), 1751–9. http://doi.org/10.1162/08997660360675026

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014a). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, *10*(11), 1–29. http://doi.org/10.1371/journal.pcbi.1003915

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014b). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, *10*(11). http://doi.org/10.1371/journal.pcbi.1003915

Kheradpisheh, S. R., Ghodrati, M., Ganjtabesh, M., & Masquelier, T. (2016a). Deep Networks Resemble Human Feed-forward Vision in Invariant Object Recognition. *Scientific Reports*, *6*(32672), 1–24. http://doi.org/10.13140/RG.2.1.2235.0564

Kheradpisheh, S. R., Ghodrati, M., Ganjtabesh, M., & Masquelier, T. (2016b). Humans and Deep Networks Largely Agree on Which Kinds of Variation Make Object Recognition Harder. *Frontiers in Computational Neuroscience*, *10*(August), 1–15. http://doi.org/10.3389/fncom.2016.00092

Kietzmann, T. C., Gert, A., Tong, F., & König, P. (2017). Representational Dynamics of Facial Viewpoint Encoding. *Journal of Cognitive Neuroscience*, 1–15. http://doi.org/10.1162/jocn

Kietzmann, T. C., Swisher, J. D., König, P., & Tong, F. (2012). Prevalence of Selectivity for Mirror-Symmetric Views of Faces in the Ventral and Dorsal Visual Pathways. *Journal of Neuroscience*, *32*(34), 11763–11772. http://doi.org/10.1523/JNEUROSCI.0126-12.2012

Körding, K. P. K. P., Kayser, C., Einhäuser, W., & König, P. (2004). How are complex cell properties adapted to the statistics of natural stimuli? *Journal of Neurophysiology*, *91*(1), 206–212. http://doi.org/10.1152/jn.00149.2003

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modelling biological vision and brain information processing. *Annual Reviews of Vision Science*, *1*, 417–446.

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*. Elsevier Ltd. http://doi.org/10.1016/j.tics.2013.06.007

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*(November), 4. http://doi.org/10.3389/neuro.06.004.2008

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

Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016). Deep Neural Networks as a Computational Model for Human Shape Sensitivity. *PLoS Computational Biology*, *12*(4), e1004896. http://doi.org/10.1371/journal.pcbi.1004896

Kümmerer, M., Theis, L., & Bethge, M. (2014). Deep Gaze I-Boosting Saliency Prediction with Feature Maps Trained on ImageNet. *arXiv:1411.1045 [Cs, Q-Bio, Stat]*, (2014), 1–11. Retrieved from http://arxiv.org/abs/1411.1045%5Cnfiles/1004/arXiv-Kummerer_et_al-2014-Deep_Gaze_I-Boosting_Saliency_Prediction_with_Feature_Maps_Trained_on_ImageNet.pdf

Lange, S., & Riedmiller, M. (2010). Deep Auto-Encoder Neural Networks in Reinforcement Learning.

Lecun, Y., & Bengio, Y. (1995). Convolutional Networks for Images, Speech, and Time-Series. In *The handbook of brain theory and neural networks* (pp. 255–258). http://doi.org/10.1017/CBO9781107415324.004

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. http://doi.org/10.1038/nature14539

Lee, D. H., Zhang, S., Fischer, A., & Bengio, Y. (2015). Difference target propagation. In *Lecture Notes in Computer Science* (Vol. 9284, pp. 498–515). http://doi.org/10.1007/978-3-319-23528-8_31

Leibo, J. Z., Liao, Q., Freiwald, W., Anselmi, F., & Poggio, T. (2017). View-tolerant face recognition and Hebbian learning imply mirror-symmetric neural tuning to head orientation. *Current Biology*, *27*, 62–67. Retrieved from http://arxiv.org/abs/1606.01552

Levy, I., Hasson, U., Avidan, G., Hendler, T., & Malach, R. (2001). Center–periphery organization of human object areas. *Nature Neuroscience*, *4*(5), 533–539. Retrieved from http://www.nature.com/neuro/journal/v4/n5/abs/nn0501_533.html

Li, N., & DiCarlo, J. J. (2008). Unsupervised natural experience rapidly alters invariant object representation in visual cortex. *Science (New York, N.Y.)*, *321*(5895), 1502–7. http://doi.org/10.1126/science.1160028

Li, N., & DiCarlo, J. J. (2010). Unsupervised Natural Visual Experience Rapidly Reshapes Size-Invariant Object Representation in Inferior Temporal Cortex.

Neuron, *67*(6), 1062–1075. http://doi.org/10.1016/j.neuron.2010.08.029

Li, Z., Yang, Y., Liu, X., Wen, S., & Xu, W. (2017). Dynamic Computational Time for Visual Attention. *arXiv Preprint*. Retrieved from http://arxiv.org/abs/1703.10332

Liang, M., & Hu, X. (2015). Recurrent convolutional neural network for object recognition. *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, (Figure 1), 3367–3375. http://doi.org/10.1109/CVPR.2015.7298958

Liao, Q., & Poggio, T. (2016). Bridging the Gaps Between Residual Learning, Recurrent Neural Networks and Visual Cortex. *arXiv Preprint*, (47), 1–16. Retrieved from http://arxiv.org/abs/1604.03640

Lillicrap, T. P., Cownden, D., Tweed, D. B., Akerman, C. J., Bell, C., Bodznick, D., … Bengio, Y. (2016). Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications*, *7*, 13276. http://doi.org/10.1038/ncomms13276

Majaj, N. J., Hong, H., Solomon, E. A., & DiCarlo, J. J. (2015). Simple Learned Weighted Sums of Inferior Temporal Neuronal Firing Rates Accurately Predict Human Core Object Recognition Performance. *Journal of Neuroscience*, *35*(39), 13402–13418. http://doi.org/10.1523/JNEUROSCI.5181-14.2015

Marblestone, A. H., Wayne, G., & Kording, K. P. (2016). Towards an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience*, *10*, 1–41. http://doi.org/10.1101/058545

Matsumoto, N., Okada, M., Sugase-Miyamoto, Y., Yamane, S., & Kawano, K. (2005). Population dynamics of face-responsive neurons in the inferior temporal cortex. *Cerebral Cortex*, *15*(8), 1103–1112. http://doi.org/10.1093/cercor/bhh209

Mitchell, T. M., Shinkareva, S. V, Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science (New York, N.Y.)*, *320*(5880), 1191–5. http://doi.org/10.1126/science.1152876

Mnih, V., Heess, N., Graves, A., & Kavukcuoglu, K. (2014). Recurrent Models of Visual Attention. *Advances in Neural Information Processing Systems 27*. http://doi.org/ng

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. a, Veness, J., Bellemare, M. G., … Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533. http://doi.org/10.1038/nature14236

Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P. A., & Kriegeskorte, N. (2013). Human object-similarity judgments reflect and transcend the primate-IT object representation. *Frontiers in Psychology*, *4*(MAR), 1–22. http://doi.org/10.3389/fpsyg.2013.00128

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, *56*(2), 400–410. http://doi.org/10.1016/j.neuroimage.2010.07.073

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., & Gallant, J. L. (2009). Bayesian Reconstruction of Natural Images from Human Brain Activity. *Neuron*, *63*, 902–915.

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep Neural Networks are Easily Fooled. *Computer Vision and Pattern Recognition, 2015 IEEE Conference on*, 427–436. http://doi.org/10.1109/CVPR.2015.7298640

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N.

(2014). A Toolbox for Representational Similarity Analysis. *PLoS Computational Biology*, *10*(4). http://doi.org/10.1371/journal.pcbi.1003553

O'Reilly, R. C. (1996). Biologically Plausible Error-driven Learning using Local Activation Differences: The Generalized Recirculation Algorithm. *Neural Computation*, *8*(5), 895–938. http://doi.org/10.1162/neco.1996.8.5.895

Olshausen, B., & Field, D. J. D. J. (2005). How close are we to understanding v1? *Neural Computation*, *17*(8), 1665–1699. http://doi.org/10.1162/0899766054026639

Olshausen, B., & Field, D. J. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(13), 607–609. http://doi.org/10.1038/381607a0

Orban, G., Berkes, P., Fiser, J., & Lengyel, M. (2016). Neural Variability and Sampling-Based Probabilistic Representations in the Visual Cortex. *Neuron*, *92*(2), 530–543. http://doi.org/10.1016/j.neuron.2016.09.038

Reichert, D. P., & Serre, T. (2013). Neuronal Synchrony in Complex-Valued Deep Networks. *International Conference on Learning Representations*. Retrieved from http://arxiv.org/abs/1312.6115

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–25. http://doi.org/10.1038/14819

Roelfsema, P. R., Lamme, V. A., & Spekreijse, H. (1998). Object-based attention in the primary visual cortex of the macaque monkey. *Nature*, *395*(6700), 376–81. http://doi.org/10.1038/nature02119.1.

Rolls, E. T. (2012). Invariant Visual Object and Face Recognition: Neural and Computational Bases, and a Model, VisNet. *Frontiers in Computational Neuroscience*, *6*(June), 35. http://doi.org/10.3389/fncom.2012.00035

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*. http://doi.org/10.1038/323533a0

Sak, H., Senior, A., & Beaufays, F. (2014). Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition. *arXiv Preprint arXiv:1402.1128*, (Cd).

Siegel, M., Donner, T., & Engel, A. (2012). Spectral fingerprints of large-scale neuronal interactions. *Nature Reviews Neuroscience*, *13*(February), 20–25. http://doi.org/10.1038/nrn3137

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural Image Statistics And Neural Representation. *Annual Review of Neuroscience*, *24*, 1193--216.

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv Preprint*, *1409.15506*, 1–14. http://doi.org/10.1016/j.infsof.2008.09.005

Spoerer, C. J., McClure, P., & Kriegeskorte, N. (2017). Recurrent convolutional neural networks: a better model of biological object recognition under occlusion. *bioRxiv*.

Srivastava, R. K., Greff, K., & Schmidhuber, J. (2015). Highway Networks. *arXiv Preprint*. Retrieved from http://arxiv.org/abs/1505.00387%5Cnhttp://www.arxiv.org/pdf/1505.00387.pdf

Sugase, Y., Yamane, S., Ueno, S., & Kawano, K. (1999). Global and fine information coded by single neurons in the temporal visual cortex. *Nature*, *400*(6747), 869–73. http://doi.org/10.1038/23703

Sun, Y., Wang, X., & Tang, X. (2015). Deeply learned face representations are sparse,

selective, and robust. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on* (Vol. 07–12–June, pp. 2892–2900). http://doi.org/10.1109/CVPR.2015.7298907

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *NIPS*, 1–9. http://doi.org/10.1007/s10107-014-0839-0

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *07–12–June*, 1–9. http://doi.org/10.1109/CVPR.2015.7298594

Taigman, Y., Ranzato, M. A., Aviv, T., & Park, M. (2014). DeepFace: Closing the Gap to Human-Level Performance in Face Verification. *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. http://doi.org/10.1109/CVPR.2014.220

Tavanaei, A., & Maida, A. S. (2016). Bio-Inspired Spiking Convolutional Neural Network using Layer-wise Sparse Coding and STDP Learning. *arXiv Preprint*, (1611.03000v2), 1–20. Retrieved from http://arxiv.org/abs/1611.03000

Tsao, D. Y., Moeller, S., & Freiwald, W. A. (2008). Comparing face patch systems in macaques and humans. *Proceedings of the National Academy of Sciences*, *105*(49), 19514. Retrieved from http://www.pnas.org/content/105/49/19514.short

Uetz, R., & Behnke, S. (2009). Locally-connected hierarchical neural networks for gpu-accelerated object recognition. *NIPS 2009 Workshop on Large-Scale Machine Learning: Parallelism and Massive Datasets*, (December), 10–13.

VanRullen, R. (2017). Perception Science in the Age of Deep Neural Networks. *Frontiers in Psychology*, *8*(February), 142. http://doi.org/10.3389/fpsyg.2017.00142

Wallis, G., & Bülthoff, H. H. (2001). Effects of temporal association on recognition memory. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(8), 4800–4. http://doi.org/10.1073/pnas.071028598

Wallis, G., & Rolls, E. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology*, *51*, 167–194. Retrieved from http://www.sciencedirect.com/science/article/pii/S0301008296000548

Wallis, T. S. A., Bethge, M., & Wichmann, F. A. (2016). Testing models of peripheral encoding using metamerism in an oddity paradigm. *Journal of Vision*, *16*(2), 1–30. http://doi.org/10.1167/16.2.4.

Walther, D. B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2009). Natural Scene Categories Revealed in Distributed Patterns of Activity in the Human Brain. *Journal of Neuroscience*, *29*(34), 10573–10581. http://doi.org/10.1523/JNEUROSCI.0559-09.2009

Weiller, D., Märtin, R., Dähne, S., Engel, A. K., & König, P. (2010). Involving motor capabilities in the formation of sensory space representations. *PloS One*, *5*(4), e10377. http://doi.org/10.1371/journal.pone.0010377

Whittington, J. C. R., & Bogacz, R. (2015). An approximation of the error back-propagation algorithm in a predictive coding network with local Hebbian synaptic plasticity. *arXiv Preprint*. http://doi.org/10.1162/NECO_a_00949

Wiskott, L., & Sejnowski, T. J. T. J. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, *14*(4), 715–770. http://doi.org/10.1162/089976602317318938

Wu, M. C.-K., David, S. V., & Gallant, J. L. (2006). Complete Functional Characterization of Sensory Neurons By System Identification. *Annual Review of Neuroscience*, *29*(1), 477–505. http://doi.org/10.1146/annurev.neuro.29.051605.113024

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., … Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv Preprint*, *1609.08144*, 1–23. Retrieved from http://arxiv.org/abs/1609.08144

Wyatte, D., Curran, T., & O'Reilly, R. (2012). The Limits of Feedforward Vision: Recurrent Processing Promotes Robust Object Recognition when Objects Are Degraded. *Journal of Cognitive Neuroscience*, *24*(11), 2248–2261. http://doi.org/10.1162/jocn_a_00282

Wyatte, D., Jilk, D. J., & O'Reilly, R. C. (2014). Early recurrent feedback facilitates visual object recognition under challenging conditions. *Frontiers in Psychology*, *5*(JUL). http://doi.org/10.3389/fpsyg.2014.00674

Wyss, R., König, P., & Verschure, P. F. M. J. (2006). A model of the ventral visual system based on temporal stability and local memory. *PLoS Biology*, *4*(5), 836–843. http://doi.org/10.1371/Citation

Xie, X., & Seung, H. S. (2003). Equivalence of backpropagation and contrastive Hebbian learning in a layered network. *Neural Computation*, *15*(2), 441–454. http://doi.org/10.1162/089976603762552988

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365. http://doi.org/10.1038/nn.4244

Yamins, D. L., Hong, H., Cadieu, C., Solomon, E., Seibert, D., & DiCarlo, 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. http://doi.org/10.1073/pnas.1403112111

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding Neural Networks Through Deep Visualization. *International Conference on Machine Learning - Deep Learning Workshop 2015*. Retrieved from http://arxiv.org/abs/1506.06579

Zeiler, M., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *Computer Vision–ECCV 2014*, *8689*, 818–833. http://doi.org/10.1007/978-3-319-10590-1_53