Engineering a Less Artificial Intelligence

Fabian H. Sinz,[1,2,7,*] Xaq Pitkow,[6,7,8] Jacob Reimer,[6,7] Matthias Bethge,[2,3,4,5,7] and Andreas S. Tolias[6,7,8,*]
[1]Institute Bioinformatics and Medical Informatics (IBMI), University of Tübingen, Germany
[2]Bernstein Center for Computational Neuroscience, University of Tübingen, Germany
[3]Centre for Integrative Neuroscience, University of Tübingen, Germany
[4]Institute for Theoretical Physics, University of Tübingen, Germany
[5]Max Planck Institute for Biological Cybernetics, Tübingen, Germany
[6]Department of Neuroscience, Baylor College of Medicine, Houston, TX, USA
[7]Center for Neuroscience and Artificial Intelligence, BCM, Houston, TX, USA
[8]Department of Electrical and Computer Engineering, Rice University, Houston, TX, USA
*Correspondence: fabian.sinz@uni-tuebingen.de (F.H.S.), astolias@bcm.edu (A.S.T.)
https://doi.org/10.1016/j.neuron.2019.08.034

Despite enormous progress in machine learning, artificial neural networks still lag behind brains in their ability to generalize to new situations. Given identical training data, differences in generalization are caused by many defining features of a learning algorithm, such as network architecture and learning rule. Their joint effect, called "inductive bias," determines how well any learning algorithm—or brain—generalizes: robust generalization needs good inductive biases. Artificial networks use rather nonspecific biases and often latch onto patterns that are only informative about the statistics of the training data but may not generalize to different scenarios. Brains, on the other hand, generalize across comparatively drastic changes in the sensory input all the time. We highlight some shortcomings of state-of-the-art learning algorithms compared to biological brains and discuss several ideas about how neuroscience can guide the quest for better inductive biases by providing useful constraints on representations and network architecture.

1. Introduction

The brain is an intricate system distinguished by its ability to learn to perform complex computations underlying perception, cognition, and motor control—defining features of intelligent behavior. For decades, scientists have attempted to mimic its abilities in artificial intelligence (AI) systems. These attempts had limited success until recent years when successful AI applications have come to pervade many aspects of our everyday life. Machine learning algorithms can now recognize objects and speech and have mastered games like chess and Go, even surpassing human performance (i.e., DeepMind's AlphaGo Zero). AI systems promise an even more significant change to come: improving medical diagnoses, finding new cures for diseases, making scientific discoveries, predicting financial markets and geopolitical trends, and identifying useful patterns in many other kinds of data.

Our perception of what constitutes intelligent behavior and how we measure it has shifted over the years as tasks that were considered hallmarks of human intelligence were solved by computers while tasks that appear to be trivial for humans and animals alike remained unsolved. Classical symbolic AI focused on reasoning with rules defined by experts, with little or no learning involved. The rule-based system of Deep Blue, which defeated Kasparov in 1997 in chess, was entirely determined by the team of experts who programmed it. Unfortunately, it did not generalize well to other tasks. This failure and the challenge of artificial intelligence even today are summarized in "Moravec's paradox" (Moravec, 1988): "it is comparatively easy to make computers exhibit adult level performance on intelligence tests or playing checkers, and difficult or impossible

to give them the skills of a one-year-old when it comes to perception and mobility." While rules in symbolic AI provide a lot of structure for generalization in very narrowly defined tasks, we find ourselves unable to define rules for everyday tasks—tasks that seem trivial because biological intelligence performs so effortlessly well.

The renaissance of AI is a result of a major shift of methods from classical symbolic AI to connectionist models used by machine learning. The critical difference from rule-based AI is that connectionist models are "trained," not "programmed." Searching through the space of possible combinations of rules in symbolic AI is replaced by adapting parameters of a flexible nonlinear function using optimization of an objective (goal) that depends on data. In artificial neuronal networks, this optimization is usually implemented by backpropagation, an algorithm developed by Paul Werbos in his PhD thesis in 1974 (Werbos, 1974). A considerable amount of effort in machine learning is being devoted to figuring out how this training can be done most effectively, as judged by how well the learned concepts generalize and how many data points are needed to robustly learn a new concept ("sample complexity").

The current state-of-the-art methods in machine learning are dominated by deep learning: multi-layer (deep) artificial neural networks (DNNs, Figure 1), which draw inspiration from the brain. Most fundamental is the idea of neurons as elementary adaptive nonlinear processing units (McCulloch and Pitts, 1943), which includes the notion of analog computation that is not well captured by the toolbox of formal logic (Rosenblatt, 1957). Each artificial neuron aggregates inputs from other neurons using weighted summation analogous to synaptic weights
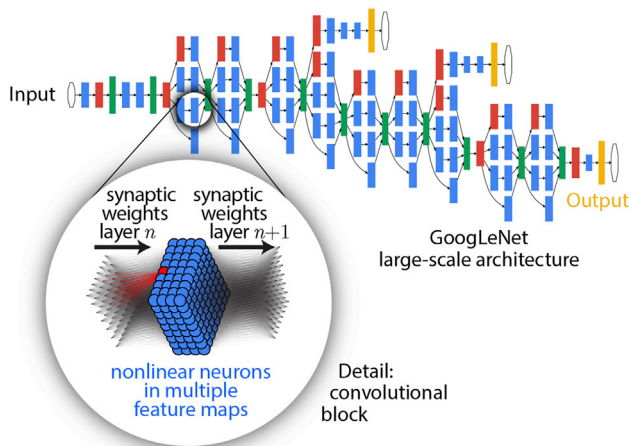
Figure 1. General Structure of Artificial Deep Neural Networks
Shown is the large-scale feedforward convolutional network architecture of GoogLeNet (Szegedy et al., 2015). Detail illustrates the smaller-scale structure within each layer, which comprises a set of feature maps and their input and output synaptic connections. Neurons in these layers typically tile visual space with spatially shifted copies of the same input and output weights (one example pattern of input weights is shown in red). For visual processing, this produces a three-dimensional array of neurons: length × width × features. These neurons apply a simple nonlinear function to their pooled inputs.

of real neurons, followed by a simple nonlinearity such as a rectifier (ReLU) or a sigmoid function (logistic function or tanh) analogous to input-output nonlinearities of neurons. Deep networks arrange their neurons in several layers, where each layer provides the input to the neurons in the next layer, analogous to the multitude of hierarchically organized brain areas for processing visual information, for example. In some deep learning architectures, local competition implemented by a winner-take-all operation (max pooling) is reminiscent of local competitive inhibitory interactions in brain circuits. Despite these similarities, the elements of artificial neural networks strongly abstract from neurophysiological details. In convolutional networks, the linear summation coefficients are shared across space (i.e., there is a neuron with exactly the same linear receptive field at each spatial location), and during learning, weights change for all locations at once. This massively reduces the number of parameters that need to be learned from data. All neurons with the same receptive field shape (but shifted to different locations) are assembled into a "feature channel," and there can be many feature channels per layer in a neuronal network. A lot of these ingredients have been around for several decades already, but thanks to a combination of training on very large datasets, advances in computing hardware, the development of software libraries, and a lot of tuning of the training schemes, it is now possible to train very large neural networks.

While machine learning had been studied only by a small crowd of academic researchers up until this decade, the success of deep learning in solving real-world problems has generated massive interest from industry and led to a complete paradigm shift in the field. Within a few years, machine learning has become *the* key technology used in virtually all AI applications. Importantly, the same learning approach that enabled AlphaGo

to achieve superhuman performance in Go has also been used to learn other games like shogi or even win against some of the best chess programs like Stockfish. Because of the ability of this approach to generalize, it represents a much more profound leap in intelligence than Deep Blue.

With the help of deep networks, it is now possible to solve some perceptual tasks that are simple for humans but used to be very challenging for AI. The so-called ImageNet benchmark (Russakovsky et al., 2015), a classification task with 1,000 categories on photographic images downloaded from the internet, played an important role in demonstrating this. Besides solving this particular task at human-level performance (He et al., 2016), it also turned out that pre-training deep networks on ImageNet can often be surprisingly beneficial for all kinds of other tasks (Donahue et al., 2014). In this approach, called "transfer learning," a network trained on one task, such as object recognition, is reused in another task by removing the task-specific part (layers high up in the hierarchy) and keeping the nonlinear features computed by the hidden layers of the network. This makes it possible to solve tasks with complex deep networks that usually would not have had enough training data to train the network *de novo*. In many computer vision tasks, this approach works much better than handcrafted features that used to be state of the art for decades. In saliency prediction, for example, the use of pre-trained features has led to a dramatic improvement of the state of the art (Kümmerer et al., 2015, 2018). Similarly, transfer learning has proven extremely useful in the behavioral tracking of animals: using a pre-trained network and a small number of training images (≈200) for fine-tuning enables the resulting network to perform very close to human-level labeling accuracy (Mathis et al., 2018; Insafutdinov et al., 2016; Pishchulin et al., 2016).

As recently pointed out by Sutton (2019), the bitter lesson of AI is that flexible methods so far have always outperformed handcrafted domain knowledge in the long run. Search-based methods of Deep Blue beat strategies attempting a deeper analytic understanding of the game, and DNNs consistently outperform handcrafted features used for decades in computer vision. However, flexibility alone cannot be the silver bullet. Without the right (implicit) assumptions, generalization is impossible (Mitchell, 1980; Wolpert and Macready, 1995, 1997). While the success of deep networks on narrowly defined perceptual tasks is a major leap forward, the range of generalization of these networks is still limited. The major challenge in building the next generation of intelligent systems is to find sources for good implicit biases that will allow for strong generalization across varying data distributions and rapid learning of new tasks without forgetting previous ones. These biases will need to be problem domain specific. Because biological brains excel at so many relevant real-world problems, it is worthwhile to ponder how they can be used as a source for good inductive biases. In the following, we lay out a few insights and ideas in this direction.

2. Current Limits of AI
The impressive—sometimes superhuman—performance of DNNs in many complex perceptual tasks might suggest that their sensory representations and decision making are similar to humans. Indeed, there seems to be an overlap between the sensory representations that DNNs trained on object recognition
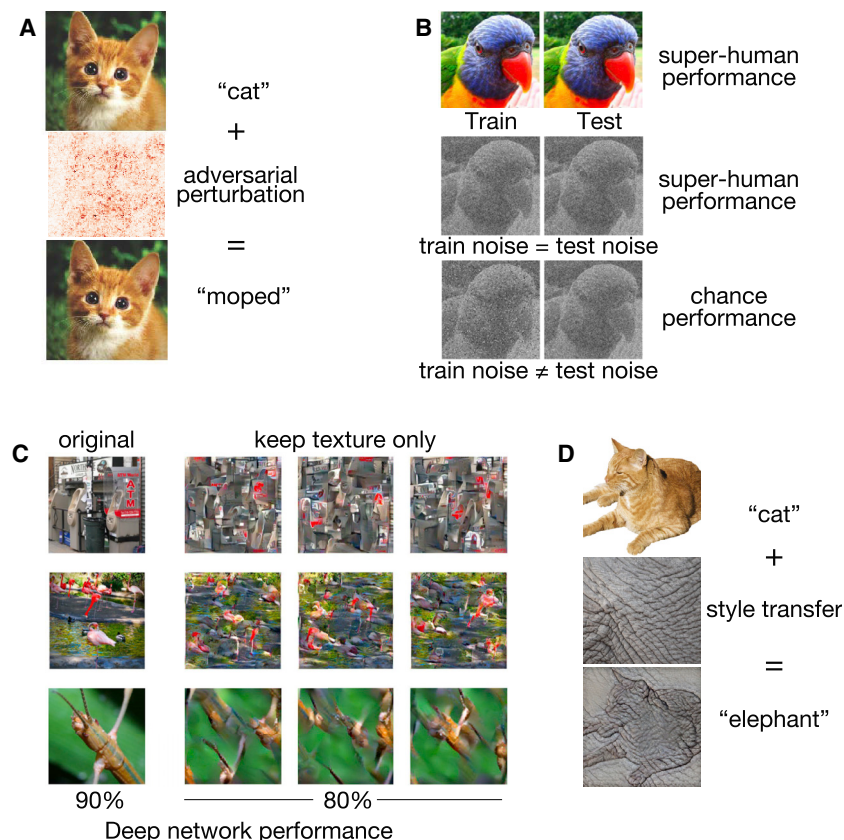
Figure 2. Despite Impressive Successes, Deep Neural Networks Have Many Important Failure Modes

(A) While DNNs trained on object recognition reach almost human-like performance on clean images (top), there exist minimal perturbations (center) that, if added to the image, can completely derail their prediction (bottom). Perceptually, humans can see almost no difference between the clean and the perturbed image even with close inspection (Szegedy et al., 2013).

(B) When networks are trained on standard color images and tested on color images, they outperform humans (top). Similarly, when trained and tested on images with the same type of noise, the performance is superhuman (middle). However, when tested on a different type of noise than at training time, the performance is at chance level (bottom). Human observers have no trouble classifying the images correctly (Geirhos et al., 2018).

(C) Examples of original and texturized images using neuronal style transfer. A vanilla VGG-16 still reaches high accuracy on the texturized images while humans suffer greatly from the loss of global shapes in many images (Brendel and Bethge, 2019).

(D) Deep networks have a texture bias. When the shape and the texture of a class are put in conflict, deep networks tend to decide based on the texture while humans decide based on the shape (Geirhos et al., 2019).

tasks create and representations measured in primate brains (Yamins et al., 2014; Yamins and DiCarlo, 2016; Khaligh-Razavi and Kriegeskorte, 2014; Güçlü and van Gerven, 2014; Cadena et al., 2019). However, even though DNNs perform well when the conditions at training and test time do not differ too much, testing them outside of their training domain demonstrates that the nature of generalization and decision making is qualitatively different from biological sensory systems.

2.1 Lack of Robustness against Changes in the Input Statistics: Adversarial Examples and Domain Adaptation

Humans have impressive generalization capabilities, and behavioral neuroscience studies suggest that the ability to generalize categories and rules to novel situations and stimuli is also present in many other animals, including rodents (Murphy et al., 2008), birds (Vaughan, 1988; Soto and Wasserman, 2014), and monkeys (Minamimoto et al., 2010). The exact meaning of "generalization beyond the training set" is much harder to define for animals that have had a lifetime of diverse visual experience with natural scene statistics (Tenenbaum et al., 2011) and generations of ancestors that were selected through evolutionary pressure to have a good architecture in that environment (Zador, 2019). Nonetheless, it is clear that artificial networks lack several key generalization capabilities compared to biological brains.

A particularly striking example of the gap between humans and machines are "minimal adversarial perturbations" of the input image discovered in computer vision networks (Szegedy et al., 2013). Adversarial perturbations are virtually imperceptible

to humans but can flip the prediction of DNNs to any desired target class (Figure 2A). This means that the decision boundaries of all classes are extremely close to any given input sample. To the best of our current knowledge, this is not the case for humans under normal viewing conditions (one study finds a small effect under time limited viewing conditions [Elsayed et al., 2018]) and highlights that DNNs lack human-level scene understanding and do not rely on the same causal features as humans for visual perception.

One key problem in making networks less vulnerable to adversarial examples is the difficulty of reliably evaluating model robustness. It has been repeatedly shown (Athalye et al., 2018; Athalye and Carlini, 2018) that virtually all defenses against adversarial examples proposed in the literature do not increase model robustness per se but merely prevent existing attacks from properly finding minimal adversarial examples. Until recently, the only defense considered effective (Athalye et al., 2018) was a particular type of training explicitly designed to guard against adversarial attacks (Madry et al., 2018). However, a recent paper (Schott et al., 2019) showed that the defending network does not learn more causal, human-like features but instead just exploits the binary nature of the dataset (MNIST, a collection of handwritten digits) and is thus unlikely to generalize to all natural images. Thus, current networks are not robust to adversarial examples, even on the simplest toy datasets of machine learning, such as MNIST. Understanding why the only existing robust systems—biological visual systems—are not vulnerable to adversarial computations could be an important guidance to the next generation of DNNs.

Domain adaptation is another striking example of the difference in generalization between biological and artificial vision systems and an opportunity for benchmarking the robustness of machine learning algorithms with direct practical relevance (Wang and Deng, 2018; Donahue et al., 2014). Humans generalize across a wide variety of changes in the input distribution, such as vast differences in illumination, changing scene contexts, or image distortions from snowflakes or rain. While humans are certainly exposed to a number of such distortions during their lifetime, there seems to be a fundamental difference in how the visual system generalizes to new inputs from a distribution that has not been previously experienced. The ability to generalize beyond the standard assumption of independent and identically distributed (i.i.d.) samples at test time would be highly desirable for machine learning algorithms, as many real-world applications involve such shifts in the input distribution. For instance, recognition systems of autonomous driving cars should be robust against a large spectrum of weather phenomena that they might not have experienced at training time, such as ash falling from a nearby volcano. Thus, the general robustness against input distortions by different types of noise can be used as one relevant case study to test the generalization beyond the i.i.d. assumption in machines and humans.

A recent study by Geirhos et al. (2018) demonstrated that humans generalize much better across different image distortions than deep networks, even though deep networks perform well on the distortions if they had access to them at training time (Figure 2B). The study explored the effect of twelve different types of low-level noise on the object recognition performance of both humans and machines. The latter were trained on either clean or distorted images (Geirhos et al., 2018). When the networks were tested on the same domain on which they were trained (i.e., the same type of noise), they consistently outperformed human observers by a large margin, showing that the networks were able to "solve" the distortions under i.i.d. train-test conditions. However, when the noise distribution during testing differed from noise seen during training, the performance of the networks was very low even for small distortions.

2.2 Decision Making in Deep Networks

The lack of robustness to simple changes in the input statistics indicates that deep networks lack human-like scene understanding. In particular, they seem to lack integration of long-range dependencies between elements within images, such as different parts of an object.

A recent study tested this hypothesis by probing deep networks for the kind of information used in decision making and found that they mostly rely on local features and largely ignore their spatial arrangement (Brendel and Bethge, 2019). The key approach in this study was to build a network that had particular properties by design and to subsequently demonstrate that it behaves very similar to standard deep architecture. To this end, the authors designed networks in which neurons in the last convolutional layer only looked at very small patches in the input image. The activity of the final layer was subsequently summed across space before it was fed into a linear classifier for object recognition. By construction this network is invariant against the exact position of a particular patch in the image, which is why it was named the "Bag-of-Feature" (BoF) network.

In addition to this invariance property, the design of the network also allowed the authors to quantify how much each image patch contributes to the decision of the network.

Brendel and Bethge (2019) compared the behavior of this BoF model to VGG-16, which is a widely used architecture in computer vision. First, they established that their BoF model can achieve a comparable performance to VGG-16 (Simonyan and Zisserman, 2015). They subsequently showed that a number of key features of linear BoF models also hold true for VGG-16. First, the BoF model predicts that shuffling local features should not affect the classification performance as the local feature patch histogram is unaffected (Figure 2C). They further corroborated this by demonstrating that the performance of VGG-16 only drops from 90.1% to 79.4% on texturized images based on neural style transfer (Figure 2D) (Gatys et al., 2015). This suggests that in stark contrast to humans, VGG-16 does not rely on global shape integration for perceptual discrimination but rather on statistical regularities in the histogram of local image features. Second, the linear classifier on top of the bag of features predicts that manipulations in separate parts of the image should not interact, which they also find to be true for VGG. Finally, they demonstrate that BoF models and VGG-16 make similar errors, and, with the help of saliency techniques, show that VGG-16 uses very similar image features for decision making as BoF models. This indicates why VGG and similar networks generalize poorly: they are extracting local features and ignoring informative large-scale structure in the input data.

3. Better Generalization through Constraints

In some sense, it is surprising that ImageNet can be solved to high accuracy with only bags of small visual words. This finding alone already suggests that DNNs trained on this task learn only the statistical regularities present in local image features, since there is no selective pressure from the objective function used during training to do otherwise. Learning to extract larger image features such as global object shapes, which are highly variable and are presented only a small number of times (number of training images per class), is much more challenging than to learn the statistical relationship between class identity and thousands of local image features present in each sample. This inductive argument, as well as the additional evidence presented above, suggests that object recognition alone is insufficient to force DNNs to learn a more physical and causal representation of the world.

The shortcomings described above suggest that the next generation of intelligent algorithms will not be achieved by following the current strategy of making networks larger or deeper. Perhaps counterintuitively, it might be the exact opposite. We know already that networks have enough capacity to express most functions because the class of networks that have only one layer of neurons with sigmoidal activation functions can theoretically fit any continuous function provided there are enough neurons (Cybenko, 1989). Even with a limited number of neurons, there is currently little evidence that deep networks are limited in their capacity to fit our current datasets. In fact, one of the first steps of practitioners is often to overfit the network on the training data to assert adequate power for a particular dataset. Similarly, the study on noise robustness by Geirhos et al. (2018) discussed above shows that networks
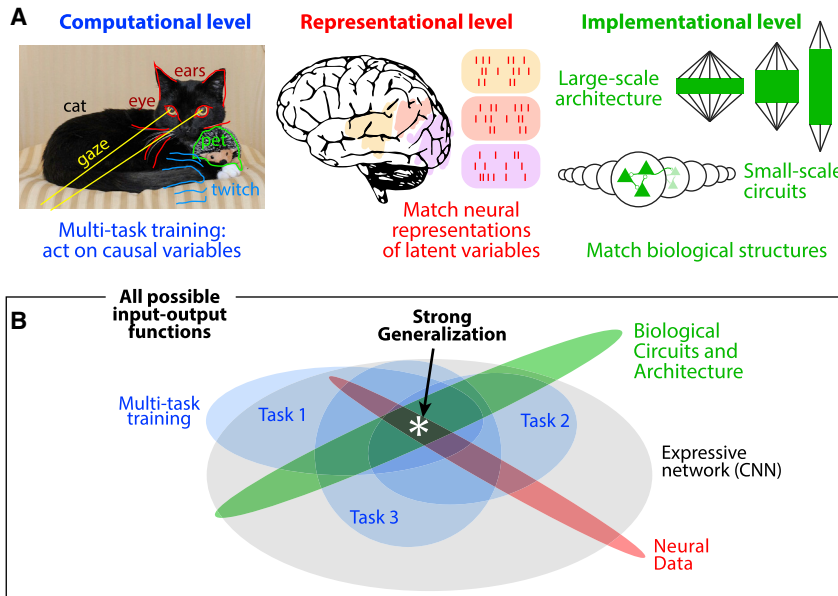
Figure 3. Improving Deep Networks at Three Levels

(A) Instead of training networks on narrow tasks like object classification, it will be better to use "multi-task training," where the network is rewarded for correct performance on diverse low- and high-level tasks involving latent variables across scales and complexity. Networks can be trained to generate latent representations that are similar to those observed in functioning brains. Finally, networks can be endowed with biological structure at the implementational level, matching architectural and/or microcircuit features. These types of improvements relate to Marr's three levels of analysis (Marr, 1982).

(B) These different levels provide complementary constraints on the space of possible solutions. Many network architectures are so expressive that they can not only learn to provide natural images with appropriate labels but can even learn to match randomly permuted labels (Zhang et al., 2016). Such networks generalize weakly within their training set but perform poorly outside of that set. Multi-task training for the same network provides additional restrictions (blue). We get additional constraints by enforcing that hidden layers in artificial networks can predict neural responses, thereby pulling representations toward those of a successful strong generalization machine, the brain (red). Finally, by constraining network structures and operations to mimic those measured in the brain, canonical operations (green), We expect that the intersection of these constraints will produce networks that have stronger generalization performance.

can be trained on each single type of noise distortion, suggesting that network capacity is not the limit. Thus, there is probably a very large group of networks, our visual system included, that can solve single tasks such as ImageNet, but they might use vastly different solution strategies and exhibit quite different robustness and generalization properties. This implies that our current datasets, even though they contain millions of examples, simply do not provide enough constraints to direct us toward a solution that is similar enough to our visual system to exhibit its desirable robustness and generalization properties. Therefore, the challenge is to come up with learning strategies that single out those well-generalizing networks among the many networks that can fit a particular dataset. One way to do that is to constrain the class of networks to narrow it down to solutions that generalize well. In other words, we need to add more bias to the class of models.

It is helpful to distinguish two types of bias, which we will call "model bias" and "inductive (or learning) bias". Model bias works like a prior probability in Bayesian inference: given some input that is inevitably ambiguous, a "fixed" network will favor certain interpretations over others or may exclude some interpretations entirely. "Inductive or learning bias" determines which fixed network is picked by the learning algorithm from the class of models given the set of training data. By "class of models," we mean a set of functions from inputs to predictions. A learning algorithm picks one function from that set of functions (also called "hypothesis space"). For instance, for a given network architecture, all networks with different values for their synaptic weights constitute a model class. Once the weights are fixed, we get a single model from that class, with its own model bias—that is, its own way of interpreting new inputs. However, the model class could be much bigger and also include models with different network architectures. Which weights are learned

(i.e., which inductive bias comes to bear) is affected by many aspects, such as the architecture, the learning rule or optimization procedure, the order in which data are presented, and the initial condition of the system. A good learning system for a particular problem will have an inductive bias that chooses networks that generalize well. Importantly, the inductive bias is ultimately problem specific. Mathematically, there is no *universal* inductive bias that works well on all problems (Wolpert and Macready, 1995, 1997). In the following, we are mainly discussing ideas for how neuroscience can be used to influence the inductive bias of artificial systems.

Biological systems can provide a source for inductive biases in several ways (Figure 3). First, biological organisms need to learn continually with the same neural network and thus critically rely on generalization across different tasks and domains (Rebuffi et al., 2017; van de Ven and Tolias, 2019). The more tasks to be solved with a single network, the fewer networks that can solve all of them and thus the stronger the resultant inductive bias on the class of models. The challenge is to define a good selection of tasks that can synergistically lead to a better bias and on which a single network can achieve a high generalization performance on all tasks (see Zamir et al., 2018 for a comparison of tasks in transfer learning). Because humans and other biological systems already solve a number of tasks with one brain, they can be a good source of inspiration to select tasks. Second, neurophysiological data provide a window into the evolved representations of a strongly generalizing network: the brain. By constraining an artificial network to match those representations (for example, by predicting the neural responses), we may bias the network toward reproducing the encoded latent variables that facilitate brain-like generalization. Third, the structure of a specific network introduces a particular inductive bias. This structure may be specified at a coarse scale, like the number,
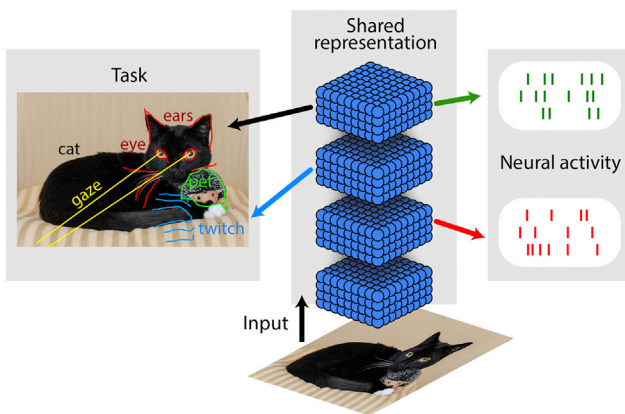
Figure 4. Neural Co-training Hypothesis
The brain has acquired a robust representation that generalizes across many tasks. Multi-task training with massive neurophysiological recordings should bias the representation of task-trained network toward these neural representations. This will not necessarily yield better performance on dataset of the task itself but could improve other aspects of the network such as a more robust generalization beyond the statistics of the training set.

size, and connectivity between hidden layers or extent of neuro-modulation, and at a fine scale, like the cell types, nonlinearities, canonical wiring rules in a local circuit, and local plasticity rules. One may also attempt to define structure at even smaller scales, looking at dendritic morphology or ion channel distribution (Markram, 2006). This level of detail is, at present, only weakly constrained by available neuroscience data, and the benefits for machine learning remain unclear, so we will not consider this finest scale here. Instead, we will focus on nonlinear input-output relationships and patterns of synaptic weights in assemblies of neurons. In the next sections, we will describe how constraints at these three levels (computational, representational, and implementational) could help create better machine learning models as well as better models of the brain.

3.1 Multi-task Learning and Data Augmentation
Convincing progress in artificial intelligence must yield models that perform well on all tasks we use to measure intelligence, including robustness to alterations in the input statistics that approximately preserve task-relevant information, such as a change in texture or different levels of noise. In principle, this does not need to happen in a single model. However, biological brains are proof that systems exist that solve all tasks with a single network. From empirical and theoretical results in machine learning, we also know that solving many tasks at once helps improve the inductive bias of a class of networks by constraining the space of solutions (Caruana, 1993; Baxter, 2000; Zhang and Yang, 2017; Ruder, 2017). As far as we know, training on a single task—albeit using a large dataset such as ImageNet—is not sufficient for this purpose (see Section 2). So one approach is to take inspiration from the perceptual and cognitive abilities of biological brains to define training regimens that enforce a network class with a better inductive bias and produce networks that generalize better. Deep learning offers an excellent framework to build such integrated models (Ruder, 2017) because its algorithms are already capable of solving single tasks so

well. Two ways of carefully choosing additional training data—"data augmentation" and "multi-task learning"—can introduce an inductive bias.

In data augmentation, an existing dataset is enriched by including more examples that have been generated from the existing dataset. Data augmentations, such as random crops or flips of the input images, are common practice in a state-of-the-art machine learning algorithm. The idea is to include the input transformation that the networks should be invariant against at training time. The problem is that we do not know which transformations make networks more robust, not to mention that the generation of the augmented datasets might generally be a hard problem. The cognitive abilities of the brain can be a great source of inspiration for defining good augmentations. For instance, a recent study used neuronal style transfer to generate images with cue conflicts between shape and texture using neuronal style transfer (Geirhos et al., 2019) and asked human subjects and deep networks to classify those images. Unsurprisingly, humans had a strong shape bias while deep networks had a strong texture bias. When they trained the networks on an augmented dataset including the texturized images, the networks were not only able to solve the original problem just as well but also became more robust against noise distortions. This is presumably because the decisions were now more based on shape than texture, and shape is more robust to the noise distortions. Just training networks on different noise-distorted images could not produce this effect (Geirhos et al., 2018). This indicates that our perception can be used to generate smart data augmentations that bias the networks in useful directions. However, data augmentation has limits since the number of possible combinations of different augmentations grows exponentially with the number of augmentation methods. Unless the different augmentations are not independent, this could become infeasible as the number of augmented data points grows too quickly.

In multi-task learning, one network is trained on different tasks at the same time. There are multiple ways to combine tasks in a network (Zhang and Yang, 2017; Ruder, 2017; Caruana, 1997). The most widely used is to share features within a network. Multi-task learning theoretically (Baxter, 2000) and practically (Caruana, 1997) improves the generalization (inductive bias) and the data efficiency (Zamir et al., 2018) (the number of data points needed to reach a certain performance) for a single task. Clearly, not all tasks should be equally useful for better generalization. Ideally, they should be related so that they can profit from shared features. Here again, the brain can be a source of inspiration for tasks; for example, in order to answer the question of which network features should be shared between which tasks, it could be fruitful to look at the functional modularity of biological networks and stimulus representations across different parts of the brain.

3.2. Representations from Neuronal Data
It is not yet clear whether tasks, even if there are many of them, are sufficient to narrow down the model class, nor what features in a network should be shared between which tasks. A stronger hypothesis is that constraining the network class with the intermediate representations of brains that can achieve the desired performance will bias the network class even further in the right direction (Figure 4). This section discusses this hypothesis,

which could not only lead to better AI models but also a better understanding of our brain by linking representations and tasks in a single model.

Past attempts of including structure from neuroscience were focused on first-order properties, such as Gabor-shaped receptive fields in early layers, or other functional properties derived from parametric models of neuronal activity (Riesenhuber and Poggio, 1999). However, these early attempts might have been too constrained and did not perform as well as more flexible deep convolutional networks. One central question is: to modify these flexible networks to avoid their shortcomings (see Section 2), *which* properties should we try to transfer from neuronal systems? One possible approach is to use a single network to solve two tasks, one being the practical task of interest and the other being to predict neuronal responses.

The task of predicting neuronal activity **r** from input stimuli **z** is referred to as "system identification" and is the most immediate and quantifiable way to capture the dynamics of activity underlying neuronal computations. System identification has a long tradition in neuroscience, especially for sensory areas such as audition (Knudsen and Konishi, 1978; Theunissen et al., 2001; Calabrese et al., 2011), touch (Chagas et al., 2013), electrosensation (Wessel et al., 1996; Gabbiani et al., 1996), olfaction (Geffen et al., 2009), and vision (Marmarelis and Naka, 1972; Chichilnisky, 2001; Pillow et al., 2008). Most neuronal system identification models can be subsumed by the framework of a generalized linear model (GLM)

$$\widehat{\mathbf{r}}(\mathbf{z}) = g(\mathbf{w}^\top \Phi(\mathbf{z})),$$

which predicts responses $\widehat{\mathbf{r}}$ through four major parts: a set of (potentially nonlinear) stimulus features $\Phi(\mathbf{z})$, a vector **w** of coefficients for combining them, a static nonlinearity g, and a loss function $\mathcal{L}(\mathbf{r}, \widehat{\mathbf{r}}, p(\mathbf{z}))$ measuring the goodness of fit between model predictions $\widehat{\mathbf{r}}(\mathbf{z})$ and biological neuronal response $\mathbf{r}(\mathbf{z})$ averaged over a distribution of inputs $p(\mathbf{z})$.

To extract computational features that are useful for more intelligent networks, we need to look beyond the simple input-output transformation of an individual neuron and extract more general features. The latent feature representation $\Phi(\mathbf{z})$ can play this role by capturing the nonlinear features required to predict neuronal responses. It characterizes all nonlinear transformations the brain has performed leading up to the responses of the target neurons. Notably, these features abstract away the biological implementation-level details of the computations.

The hypothesis that neuronal data can inform machine learning algorithms by providing additional constraints on the model class posits (1) the existence of a latent feature space that generalizes across all neurons in a certain group (e.g., the same cell type and/or the same brain area) and across stimuli and (2) that these features are useful not only for predicting neuronal data but also for performing tasks. Intuitively, (1) means that neurons in a group perform similar computations as the ones captured by the latent nonlinear representation Φ, while (2) means that these features should be useful for tasks the system might face as well. Several lines of recent work provide evidence that this might be the case.

Regarding (1), several studies demonstrated that such low-dimensional latent feature representations can be learned from data using flexible machine learning tools, such as deep networks, and that they outperform handcrafted features (McIntosh et al., 2016; Lehky et al., 1992; Lau et al., 2002; Prenger et al., 2004; Sinz et al., 2018; Cadena et al., 2019; Antolík et al., 2016; Ecker et al., 2019; Klindt et al., 2017; Zhang et al., 2019; Kindel et al., 2019; Vintch et al., 2015; Batty et al., 2016; Pandarinath et al., 2018; Walker et al., 2019; Ecker et al., 2019). When the nonlinear feature representation is shared across many neurons, it can be learned from data, even though the number of parameters in the nonlinear network might be too large to be reasonably learned from single neurons. Some of the studies simultaneously predict thousands of neurons with less than 100 hidden feature dimensions (Sinz et al., 2018; Ecker et al., 2019; Cadena et al., 2019; Klindt et al., 2017). Sinz et al. (2018) also showed that the nonlinear feature representation generalizes beyond the natural stimuli on which it was trained, successfully predicting neuronal responses to noise stimuli.

Another strong test that the features learned by deep predictive models characterize neuronal responses well is that these models can also be used to generate *new* stimuli that strongly drive the responses of biological neurons. Two recent studies independently developed a novel closed-loop experimental paradigm—called "inception loops"—combining *in vivo* recordings with *in silico* nonlinear response modeling (Walker et al., 2019; Figure 5 in Bashivan et al., 2019). The authors trained deep learning models based on shared representations Φ to accurately predict the responses of a group of neurons to natural input. Subsequently, they used these models to synthesize stimuli for driving the response of selected model neurons as strongly as possible (Figure 5A). When they showed these images back to the respective biological neurons in subsequent experiments, the neurons indeed responded more strongly to the synthesized images than to a number of control stimuli, indicating that the models capture essential elements of nonlinear neural representations.

Regarding (2), a recent line of work demonstrated that neuronal activity in monkeys and humans can also be accurately predicted when features Φ are derived from deep networks pre-trained on machine learning tasks such as object classification on ImageNet (Yamins et al., 2014; Cadieu et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014; Hong et al., 2016; Yamins and DiCarlo, 2016; Cadena et al., 2019; Güçlü and van Gerven, 2014; Cichy et al., 2016; Agrawal et al., 2014). Several of these studies also showed that there is a strong correlation between the depth of the best predicting layer in the artificial neuronal network and the depth of the predicted neuronal area in the visual hierarchy. Additionally, there is a strong correlation between how well a neural network performs on a classification task and how well some of its nonlinear features predict neuronal activity (Yamins et al., 2014; Yamins and DiCarlo, 2016). This is evidence that neuronal feature spaces Φ are related to good features for machine learning tasks.

Empirical validation of the neuronal co-training hypothesis likely needs massive amounts of neuronal data from different areas recorded during behavior. It will also be important to develop careful null hypotheses and control experiments for assessing whether neuronal data can bias deep networks toward a new generation
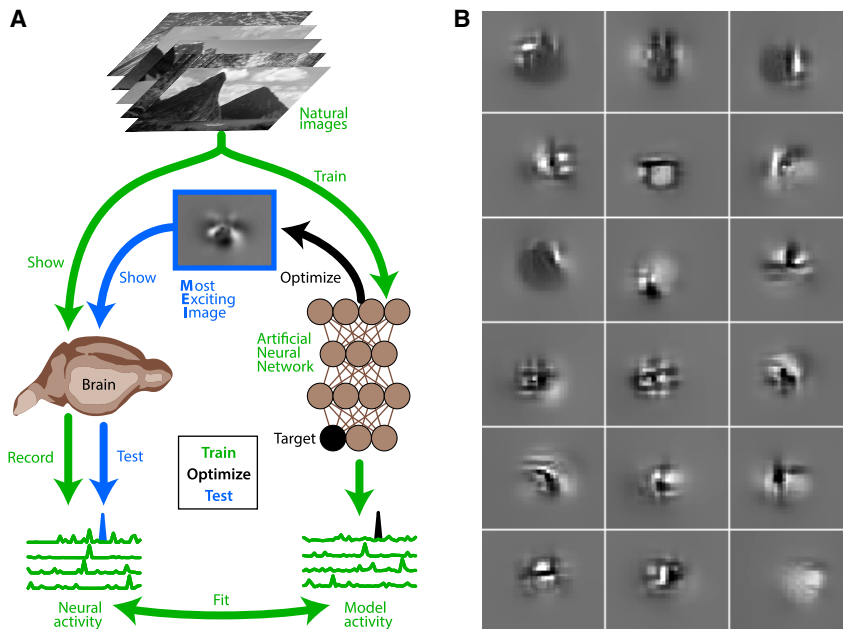
of more robust and better generalizing learning machines. Some early reports used fMRI data to constrain deep neural networks (Fong et al., 2018), but it remains to be shown whether and how co-training with neuronal data will increase the generalization of the algorithms beyond their training data domain.

3.3. Copying Structure from Biology

Another way to introduce a bias in the class of networks is through the choice of architectures (such as convolutionality), nonlinearities, and learning algorithms (Figure 6). On a macro-scale, the brain has many specialized modules, such as the hippocampus, thalamus, and basal ganglia, that have their own specialized connectivities and interactions. We currently do not know enough about all the functions of the different modules within the entire brain, and machine learning currently focuses mostly on tasks that are attributed to single modules (e.g., the cortical visual system). Therefore, the possible advantage of factorizing large networks into modules according to this network structure remains unknown, although some work has started to explore such motifs as memory modules (Weston et al., 2014; Graves et al., 2014) or elements akin to voluntary eye movements or attention (Jaderberg et al., 2015; Mnih et al., 2014; Vaswani et al., 2017). On a micro- and mesoscale, aspects of artificial neural networks have been inspired by biological networks, using such properties as normalization, winner-takes-all mechanisms like max pooling (Riesenhuber and Poggio, 1999), attention (Larochelle and Hinton, 2010), dropout (Srivastava et al., 2014), or even merely neurons as basic computational elements. However, despite this inspiration, there are also many clear differences between neuronal networks *in vivo* and *in silico*: almost every artificial neuron in the machine learning literature is modeled as a point neuron described by a scalar nonlinear function (ReLU, ELU, sigmoid, etc.) of a linear projection of its inputs. Slightly greater complexity is afforded by long short-term memory units and

gated recurrent units, which also provide gating nonlinearities (Hochreiter and Schmidhuber, 1997; Chung et al., 2014).

Real neurons are vastly more complex machines, with nonlinear interactions within and between different branches of their dendrites (London and Häusser, 2005). They thus embody their own miniature neural network within a single neuron (Poirazi et al., 2003).

Not only does the brain have more complex neural units, it has a reliably intricate circuit structure on multiple spatial scales. Cortical microcircuits comprise many genetically and functionally distinct cell types (Douglas and Martin, 1991; Jiang et al., 2015) that may perform operations like gating, homeostatic regulation, divisive normalization, and more sophisticated operations. In contrast, microcircuits in conventional neural networks use weighted sums and max pooling or pairwise multiplication and often even skip over the nonlinearities (He et al., 2016). Most artificial networks for static data use a feedforward architecture. For temporal data like speech and natural language processing, recurrent networks have been explored much more extensively, but feedforward architectures often achieve state-of-the-art performance. In contrast, the brain has a rich recurrent structured connectivity both locally within a cortical area and at the largest scale. This recurrence can be viewed as effectively making the network deeper, but with fewer parameters: the recurrent network can be unrolled to create an equivalent feedforward network with weights shared across layers. Although this unrolled network is less expressive than a network with the same architecture and depth but with untied weights, the reduced expressivity might also prove to be a useful inductive bias.

Adding more biologically plausible mechanisms into neuronal networks at the microscale would be a better model of biology, but it is unlikely that networks made from biologically plausible units or microcircuits would be limited in terms of what functions they can realize (Tripp and Eliasmith, 2016; Parisien et al., 2008): such networks may also be universal function approximators. Thus, the bias from merely using biological components alone could be weak. At the same time, changes in the architecture
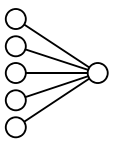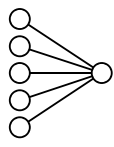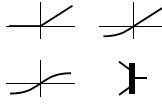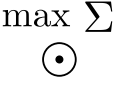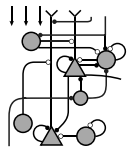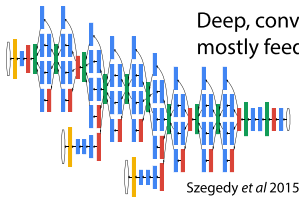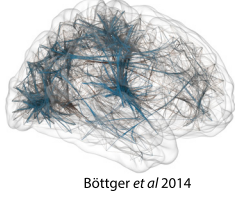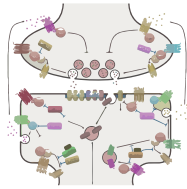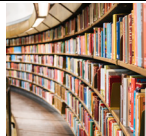
Figure 6. Core Properties of Some Traditional Artificial Neural Networks Compared to Those for Biological Neural Networks
Images from Szegedy et al. (2015), Böttger et al. (2014), and Gerber et al. (2016).

and the mechanisms of artificial neuronal networks can make a substantial difference in performance, even if the network architectures are randomly generated (Xie et al., 2019). For instance, residual networks (ResNets) (He et al., 2016) train substantially better on many tasks than standard sequential deep networks even though the only difference between them is additional shortcuts between early and later layers. But here again, the set of functions that can be realized by ResNets and sequential deep networks are very similar. The relevant difference thus lies in the interaction between loss, architecture, and learning rule.

Since the set of achievable input-output functions largely overlaps for different architectures, one can think of architectures as different parameterizations of a similar class of functions.

Learning in neural networks corresponds to moving along a trajectory in the space of network parameters. This trajectory is determined by the input data and the learning rule. In artificial neural networks, this learning rule is usually to follow the negative gradient of an objective function (Marblestone et al., 2016) using stochastic gradient descent. Changing the parameterization changes the way the learning rule interacts with the parameters. Consequently, the optimization process largely determines the inductive bias, and changes in the mechanisms or architecture must be analyzed in conjunction with the learning rule. The regularizing influence of the learning rule on the performance of neuronal networks is currently an active area of research in machine learning. Biological learning rules so far have not played an important role in training networks to high performance

(Bartunov et al., 2018). This might again be because of a mismatch between architecture and learning rule or because biological learning is so hard to study experimentally that current models of biological learning rules are inadequate.

Another consideration is that the amount of learning that occurs in the brain, in particular in early stages of sensory processing, might be limited (Zador, 2019). One line of evidence for this latter argument is that many animals already perform quite complex behaviors well, right from birth. This suggest that developmental processes, acquired through the course of evolution, provide a good initialization that can be efficiently fine-tuned with experience (Zador, 2019).

In summary, even if the architecture and elements of networks might not constrain the set of realizable functions, they do influence which function is found when learning from limited data. This means that the inductive bias of architectural elements is strongly coupled to the learning rule. Because we still know very little about biological learning in large networks or the developmental processes that initialize neuronal architectures, introducing inductive biases by building detailed biologically plausible neuronal networks seems challenging.

Instead, compact descriptions of the nonlinear feature space that most effectively identify a neural system directly reflects the model bias in the brain's neural networks. We thus propose to identify better inductive biases by *functionally* constraining artificial networks through multi-task training, using the output layers of a neural network to achieve behavioral goals while matching intermediate layers to large-scale single-cell data from neurophysiological recordings.

4. Conclusion

The history of more than 60 years of AI research is still marked by Moravec's paradox. For decades, computers have been able to outperform us in abstract but closely circumscribed situations, such as playing chess. However, tasks that our brain solves subconsciously and effortlessly, like grasping, navigation, and scene understanding, turn out to be hard to teach to machines. Brains have been evolving to accomplish feats of subconscious intelligence for much longer than to solve tasks we usually consider more difficult (e.g., cognitive behaviors depending on frontal lobes) as these feats already provide fitness for many animals. Thus, it is not surprising that this subconscious intelligence in biological systems is more difficult to match than symbolic reasoning. Building learning machines that are as flexible and versatile yet as robust and strongly generalizing as mammalian brains is the major challenge in machine learning for the next years to come. Here, we described some ways in which brains themselves could help advance AI by building new bridges between neuroscience and machine learning.

The limitation of current deep networks is not the lack of expressive power (i.e., the diversity and complexity of functions they can express). Instead, deep networks are limited because they lack the right inductive bias. Even shallow neural networks are powerful enough to express any well-behaved function (Cybenko, 1989). Although shallow universal function approximators might require a much larger number of neurons than deep networks, empirical studies suggest that they can approximate similar functions with a similar number of parameters (Ba and

Caruana, 2014). The problem is rather that training shallow networks is difficult. With increasing complexity of a network class, the number of networks that fit a given finite dataset grows as well. The challenge is then to choose a learning algorithm with a good inductive bias that selects networks out of that large set of possible candidates that generalize well to unseen data. The current success of deep networks derives from the inductive bias implicit in the combination of both the network architecture (such as convolutionality) and the learning rule based on stochastic gradient descent (backpropagation).

However, while the trained networks predict well for test samples drawn from the same distribution as training data, they use different decision strategies than humans and are much less robust to changes in input statistics that brains easily handle. To match this ability with artificial neural networks, we must improve the inductive bias of current deep networks. We discussed three possible approaches to achieve this: training each network to solve many behavioral tasks at once, co-training machine learning algorithms to match the brain's latent representations observed in neurophysiological data, and choosing a specific network architecture or weight-sharing scheme together with the right learning rule.

The last 10 years in neuroscience have witnessed a surge of new technologies that enable us to measure and analyze brain circuits in ways that we could only dream about until now. We can now record chronically from many thousands of neurons simultaneously (Sofroniew et al., 2016; Jun et al., 2017) and decipher their wiring diagram at the level of billions of synapses. The MICrONS project, funded by IARPA, is a collaboration between several institutions (http://http://www.ninai.org/) that has recorded 10^5 neurons in one mouse and used electron microscopy to measure nanoscale synaptic connectivity over a 1 mm^3 volume. We now have tools both to map the functional organization of the mammalian brain at an unprecedented level of detail and to manipulate activity with cellular localization and millisecond precision in behaving animals. In order to make sense of this deluge of data, neuroscience needs to develop new methods to link neuronal representations and architectural features to the collection of complex tasks a brain solves every day.

Deep learning can provide a framework to integrate these diverse experimental observations into one common model. Careful analysis from computational neuroscience and machine learning should continually expose the differences between biological and AI through new benchmarks, allowing us to refine the models. With experiments that probe the mechanisms of the brain's inductive biases and analyses that identify the key properties that manifest those biases, neuroscience and machine learning together can help build the next generation of artificial intelligence.

REFERENCES

Agrawal, P., Stansbury, D., Malik, J., and Gallant, J.L. (2014). Pixels to voxels: modeling visual representation in the human brain. arXiv, arXiv:1407.5104 https://arxiv.org/abs/1407.5104.

Antolík, J., Hofer, S.B., Bednar, J.A., and Mrsic-Flogel, T.D. (2016). Model constrained by visual hierarchy improves prediction of neural responses to natural scenes. PLoS Comput. Biol. 12, e1004927.

Athalye, A., and Carlini, N. (2018). On the robustness of the cvpr 2018 white-box adversarial example defenses. arXiv, arXiv:1804.03286 https://arxiv.org/abs/1804.03286.

Athalye, A., Carlini, N., and Wagner, D. (2018). Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples, In Proceedings of the 35th International Conference on Machine Learning. PMLR 80.

Ba, J., and Caruana, R. (2014). Do deep nets really need to be deep? In Advances in Neural Information Processing Systems (NIPS), pp. 2654–2662.

Bartunov, S., Santoro, A., Richards, B., Marris, L., Hinton, G.E., and Lillicrap, T. (2018). Assessing the scalability of biologically-motivated deep learning algorithms and architectures. Adv. Neural Inf. Process. Syst. 31, 9390–9400.

Bashivan, P., Kar, K., and DiCarlo, J. (2019). Neural population control via deep ANN image synthesis. In 2018 Conference on Cognitive Computational Neuroscience (CCN), pp. 1–33.

Batty, E., Merel, J., Brackbill, N., Heitman, A., Sher, A., Litke, A., Chichilnisky, E.J., and Paninski, L. (2016). Multilayer network models of primate retinal ganglion cells. In International Conference on Learning Representations https://openreview.net/forum?id=HkEI22jeg.

Baxter, J. (2000). A model of inductive bias learning. J. Artif. Intell. Res. 12, 149–198.

Böttger, J., Schäfer, A., Lohmann, G., Villringer, A., and Margulies, D.S. (2014). Three-dimensional mean-shift edge bundling for the visualization of functional connectivity in the brain. IEEE Trans. Vis. Comput. Graph. 20, 471–480.

Brendel, W., and Bethge, M. (2019). Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. In International Conference on Learning Representations https://openreview.net/forum?id=SkfMWhAqYQ.

Cadena, S.A., Denfield, G.H., Walker, E.Y., Gatys, L.A., Tolias, A.S., Bethge, M., and Ecker, A.S. (2019). Deep convolutional models improve predictions of macaque V1 responses to natural images. PLoS Comput. Biol. 15, e1006897.

Cadieu, C.F., Hong, H., Yamins, D.L.K., Pinto, N., Ardila, D., Solomon, E.A., Majaj, N.J., and DiCarlo, J.J. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. PLoS Comput. Biol. 10, e1003963.

Calabrese, A., Schumacher, J.W., Schneider, D.M., Paninski, L., and Woolley, S.M. (2011). A generalized linear model for estimating spectrotemporal receptive fields from responses to natural sounds. PLoS ONE 6, e16104.

Caruana, R.A. (1993). Multitask learning: a knowledge-based source of inductive bias. In Proceedings of the Tenth International Conference on Machine Learning (PMLR), pp. 41–48.

Caruana, R. (1997). Multitask learning. PhD thesis (Carnegie Mellon University).

Chagas, A.M., Theis, L., Sengupta, B., Stüttgen, M.C., Bethge, M., and Schwarz, C. (2013). Functional analysis of ultra high information rates conveyed by rat vibrissal primary afferents. Front. Neural Circuits 7, 190.

Chichilnisky, E.J. (2001). A simple white noise analysis of neuronal light responses. Network: Computation in Neural Systems 12, 199–213.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv, arXiv:1412.3555 https://arxiv.org/abs/1412.3555.

Cichy, R.M., Khosla, A., Pantazis, D., Torralba, A., and Oliva, A. (2016). Deep neural networks predict hierarchical spatio-temporal cortical dynamics of human visual object recognition. arXiv, arXiv:1601.02970 https://arxiv.org/abs/1601.02970.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. Math. Contr. Signals Syst. 2, 303–314.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. (2014). DeCAF: a deep convolutional activation feature for generic visual recognition. In Proceedings of the 31st International Conference on Machine Learning (PMLR), pp. 647–655.

Douglas, R.J., and Martin, K.A. (1991). A functional microcircuit for cat visual cortex. J. Physiol. 440, 735–769.

Ecker, A., Sinz, F., Froudarakis, E., Fahey, P., Cadena, S., Walker, E.Y., Cobos, E., Reimer, J., Tolias, A.S., and Bethge, M. (2019). A rotation-equivariant convolutional neural network model of primary visual cortex. In Seventh International Conference on Learning Representations (ICLR 2019), pp. 1–11.

Elsayed, G.F., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., and Sohl-Dickstein, J. (2018). Adversarial examples that fool both computer vision and time-limited humans. In Advances in Neural Information Processing Systems (NIPS 2018), pp. 3910–3920.

Fong, R.C., Scheirer, W.J., and Cox, D.D. (2018). Using human brain activity to guide machine learning. Sci. Rep. 8, 5397.

Gabbiani, F., Metzner, W., Wessel, R., and Koch, C. (1996). From stimulus encoding to feature extraction in weakly electric fish. Nature 384, 564–567.

Gatys, L.A., Ecker, A.S., and Bethge, M. (2015). A neural algorithm of artistic style. arXiv, arXiv:1508.06576 https://arxiv.org/abs/1508.06576.

Geffen, M.N., Broome, B.M., Laurent, G., and Meister, M. (2009). Neural encoding of rapidly fluctuating odors. Neuron 61, 570–586.

Geirhos, R., Temme, C.R.M., Rauber, J., Schütt, H.H., Bethge, M., and Wichmann, F.A. (2018). Generalisation in humans and deep neural networks. Adv. Neural Inf. Process. Syst. 31, 7549–7561.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., and Brendel, W. (2019). Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In International Conference on Learning Representations https://openreview.net/forum?id=Bygh9j09KX.

Gerber, K.J., Squires, K.E., and Hepler, J.R. (2016). Roles for regulator of g protein signaling proteins in synaptic signaling and plasticity. Mol. Pharmacol. 89, 273–286.

Graves, A., Wayne, G., and Danihelka, I. (2014). Neural turing machines. arXiv, arXiv:1410.5401 https://arxiv.org/abs/1410.5401.

Güçlü, U., and van Gerven, M.A.J. (2014). Deep neural networks reveal a gradient in the complexity of neural representations across the brain's ventral visual pathway. J. Neurosci. 35, 10005–10014.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (IEEE), pp. 770–778.

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. Neural Comput. 9, 1735–1780.

Hong, H., Yamins, D.L.K., Majaj, N.J., and DiCarlo, J.J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. Nat. Neurosci. 19, 613–622.

Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., and Schiele, B. (2016). Deepercut: a deeper, stronger, and faster multi-person pose estimation model. In European Conference on Computer Vision, B. Leibe, J. Matas, N. Sebe, and M. Welling, eds. (Springer), pp. 34–50.

Jaderberg, M., Simonyan, K., and Zisserman, A. (2015). Spatial transformer networks. In Advances in neural information processing systems (NIPS 2015), pp. 2017–2025.

Jiang, X., Shen, S., Cadwell, C.R., Berens, P., Sinz, F., Ecker, A.S., Patel, S., and Tolias, A.S. (2015). Principles of connectivity among morphologically defined cell types in adult neocortex. Science 350, aac9462.

Jun, J.J., Steinmetz, N.A., Siegle, J.H., Denman, D.J., Bauza, M., Barbarits, B., Lee, A.K., Anastassiou, C.A., Andrei, A., Aydın, Ç., et al. (2017). Fully integrated silicon probes for high-density recording of neural activity. Nature 551, 232–236.

Khaligh-Razavi, S.M., and Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput. Biol. 10, e1003915.

Kindel, W.F., Christensen, E.D., and Zylberberg, J. (2019). Using deep learning to probe the neural code for images in primary visual cortex. J. Vis. 19, 29.

Klindt, D., Ecker, A.S., Euler, T., and Bethge, M. (2017). Neural system identification for large populations separating "what" and "where". In Advances in Neural Information Processing Systems (NIPS 2017), pp. 3506–3516.

Knudsen, E.I., and Konishi, M. (1978). Center-surround organization of auditory receptive fields in the owl. Science 202, 778–780.

Kümmerer, M., Theis, L., and Bethge, M. (2015). Deep gaze I: boosting saliency prediction with feature maps trained on imagenet, in: ICLR Workshop. arXiv, arXiv:1411.1045 https://arxiv.org/abs/1411.1045.

Kümmerer, M., Wallis, T.S.A., and Bethge, M. (2018). Saliency benchmarking made easy: separating models, maps and metrics. In The European Conference on Computer Vision (ECCV), V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, eds. (Springer), pp. 798–814.

Larochelle, H., and Hinton, G.E. (2010). Learning to combine foveal glimpses with a third-order Boltzmann machine. Adv. Neural Inf. Process. Syst. 23, 1243–1251.

Lau, B., Stanley, G.B., and Dan, Y. (2002). Computational subunits of visual cortical neurons revealed by artificial neural networks. Proc. Natl. Acad. Sci. USA 99, 8974–8979.

Lehky, S.R., Sejnowski, T.J., and Desimone, R. (1992). Predicting responses of nonlinear neurons in monkey striate cortex to complex patterns. J. Neurosci. 12, 3568–3581.

London, M., and Häusser, M. (2005). Dendritic computation. Annu. Rev. Neurosci. 28, 503–532.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In International Conference on Learning Representations https://openreview.net/forum?id=rJzIBfZAb.

Marblestone, A.H., Wayne, G., and Kording, K.P. (2016). Toward an integration of deep learning and neuroscience. Front. Comput. Neurosci. 10, 94.

Markram, H. (2006). The blue brain project. Nat. Rev. Neurosci. 7, 153–160.

Marmarelis, P.Z., and Naka, K. (1972). White-noise analysis of a neuron chain: an application of the Wiener theory. Science 175, 1276–1278.

Marr, D. (1982). Vision: A Computational Investigation into the Human Representation and Processing of Visual Information (Henry Holt & Co.).

Mathis, A., Mamidanna, P., Cury, K.M., Abe, T., Murthy, V.N., Mathis, M.W., and Bethge, M. (2018). DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nat. Neurosci. 21, 1281–1289.

McCulloch, W.S., and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biophys. 5, 115–133.

McIntosh, L.T., Maheswaranathan, N., Nayebi, A., Ganguli, S., and Baccus, S.A. (2016). Deep learning models of the retinal response to natural scenes. Adv. Neural Inf. Process. Syst. 29, 1369–1377.

Minamimoto, T., Saunders, R.C., and Richmond, B.J. (2010). Monkeys quickly learn and generalize visual categories without lateral prefrontal cortex. Neuron 66, 501–507.

Mitchell, T.M. (1980). The need for biases in learning generalizations. Rutgers CS tech report CBM-TR-117 (Department of Computer Science, Laboratory for Computer Science Research).

Mnih, V., Heess, N., Graves, A., and Kavukcuoglu, K. (2014). Recurrent models of visual attention. Adv. Neural Inf. Process. Syst. 27, 2204–2212.

Moravec, H. (1988). Mind Children: The Future of Robot and Human Intelligence (Harvard University Press).

Murphy, R.A., Mondragón, E., and Murphy, V.A. (2008). Rule learning by rats. Science 319, 1849–1851.

Pandarinath, C., O'Shea, D.J., Collins, J., Jozefowicz, R., Stavisky, S.D., Kao, J.C., Trautmann, E.M., Kaufman, M.T., Ryu, S.I., Hochberg, L.R., et al. (2018). Inferring single-trial neural population dynamics using sequential auto-encoders. Nat. Methods 15, 805–815.

Parisien, C., Anderson, C.H., and Eliasmith, C. (2008). Solving the problem of negative synaptic weights in cortical models. Neural Comput. 20, 1473–1494.

Pillow, J.W., Shlens, J., Paninski, L., Sher, A., Litke, A.M., Chichilnisky, E.J., and Simoncelli, E.P. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. Nature 454, 995–999.

Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P.V., and Schiele, B. (2016). Deepcut: joint subset partition and labeling for multi person pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (IEEE), pp. 4929–4937.

Poirazi, P., Brannon, T., and Mel, B.W. (2003). Pyramidal neuron as two-layer neural network. Neuron 37, 989–999.

Ponce, C.R., Xiao, W., Schade, P.F., Hartmann, T.S., Kreiman, G., and Livingstone, M.S. (2019). Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. Cell 177, 999–1009.e10.

Prenger, R., Wu, M.C.K., David, S.V., and Gallant, J.L. (2004). Nonlinear V1 responses to natural scenes revealed by neural network analysis. Neural Netw. 17, 663–679.

Rebuffi, S.A., Bilen, H., and Vedaldi, A. (2017). Learning multiple visual domains with residual adapters. Adv. Neural Inf. Process. Syst. 30, 506–516.

Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. Nat. Neurosci. 2, 1019–1025.

Rosenblatt, F. (1957). The Perceptron, A Perceiving and Recognizing Automaton (Cornell Aeronautical Laboratory).

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. arXiv, arXiv:1706.05098 https://arxiv.org/abs/1706.05098.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). ImageNet Large Scale Visual Recognition Challenge. IJCV.

Schott, L., Rauber, J., Bethge, M., and Brendel, W. (2019). Towards the First Adversarially Robust Neural Network Model on MNIST. In Seventh International Conference on Learning Representations (ICLR 2019), pp. 1–16.

Simonyan, K., and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. ICLR.

Sinz, F., Ecker, A.S., Fahey, P., Walker, E., Cobos, E., Froudarakis, E., Yatsenko, D., Pitkow, Z., Reimer, J., and Tolias, A. (2018). Stimulus domain transfer in recurrent models for large scale cortical population prediction on video. Adv. Neural Inf. Process. Syst. 31, 7199–7210.

Sofroniew, N.J., Flickinger, D., King, J., and Svoboda, K. (2016). A large field of view two-photon mesoscope with subcellular resolution for in vivo imaging. eLife 5, e14472.

Soto, F.A., and Wasserman, E.A. (2014). Mechanisms of object recognition: what we have learned from pigeons. Front. Neural Circuits 8, 122.

Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. *15*, 1929–1958.

Sutton, R. (2019). The bitter lesson. https://gradientdescent.co/t/the-bitter-lesson-by-richard-sutton/187.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2013). Intriguing properties of neural networks. arXiv, arXiv:1312.6199 https://arxiv.org/abs/1312.6199.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (IEEE), pp. 1–9.

Tenenbaum, J.B., Kemp, C., Griffiths, T.L., and Goodman, N.D. (2011). How to grow a mind: statistics, structure, and abstraction. Science *331*, 1279–1285.

Theunissen, F.E., David, S.V., Singh, N.C., Hsu, A., Vinje, W.E., and Gallant, J.L. (2001). Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network: Computation in Neural Systems 12*, 289–316.

Tripp, B., and Eliasmith, C. (2016). Function approximation in inhibitory networks. Neural Netw. *77*, 95–106.

van de Ven, G.M., and Tolias, A.S. (2019). Three scenarios for continual learning. arXiv, arXiv:1904.07734 https://arxiv.org/abs/1904.07734.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. Adv. Neural Inf. Process. Syst. *30*, 5998–6008.

Vaughan, W. (1988). Formation of equivalence sets in pigeons. J. Exp. Psychol. Anim. Behav. Process. *14*, 36.

Vintch, B., Movshon, J.A., and Simoncelli, E.P. (2015). A convolutional subunit model for neuronal responses in macaque V1. J. Neurosci. *35*, 14829–14841.

Walker, E.Y., Sinz, F.H., Froudarakis, E., Fahey, P.G., Muhammad, T., Ecker, A.S., Cobos, E., Reimer, J., Pitkow, X., and Tolias, A.S. (2019). Inception in visual cortex: in vivo-silico loops reveal most exciting images. bioRxiv. https://doi.org/10.1101/506956.

Wang, M., and Deng, W. (2018). Deep visual domain adaptation: a survey. arXiv, arXiv:1802.03601 https://arxiv.org/abs/1802.03601.

Werbos, P. (1974). Beyond regression: new tools for prediction and analysis in the behavioral sciences. PhD thesis (Harvard University).

Wessel, R., Koch, C., and Gabbiani, F. (1996). Coding of time-varying electric field amplitude modulations in a wave-type electric fish. J. Neurophysiol. *75*, 2280–2293.

Weston, J., Chopra, S., and Bordes, A. (2014). Memory networks. arXiv, arXiv:1410.3916 https://arxiv.org/abs/1410.3916.

Wolpert, D.H., and Macready, W.G. (1995). No free lunch theorems for search. Technical Report. Technical Report SFI-TR-95-02-010 (Santa Fe Institute).

Wolpert, D.H., and Macready, W.G. (1997). No free lunch theorems for optimization. IEEE Trans. Evol. Comput. *1*, 67–82.

Xie, S., Kirillov, A., Girshick, R., and He, K. (2019). Exploring randomly wired neural networks for image recognition. arXiv, arXiv:1904.01569 https://arxiv.org/abs/1904.01569.

Yamins, D.L.K., and DiCarlo, J.J. (2016). Using goal-driven deep learning models to understand sensory cortex. Nat. Neurosci. *19*, 356–365.

Yamins, D.L.K., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., and DiCarlo, J.J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc. Natl. Acad. Sci. USA *111*, 8619–8624.

Zador, A.M. (2019). A critique of pure learning and what artificial neural networks can learn from animal brains. Nat. Commun. *10*, 3770.

Zamir, A., Sax, A., Shen, W., Guibas, L., Malik, J., and Savarese, S. (2018). Taskonomy: disentangling task transfer learning. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, L. O'Connor, ed. (IEEE), pp. 3712–3722.

Zhang, Y., and Yang, Q. (2017). A survey on multi-task learning. arXiv, arXiv:1707.08114 https://arxiv.org/abs/1707.08114.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2016). Understanding deep learning requires rethinking generalization, International Conference on Learning Representations. https://openreview.net/forum?id=Sy8gdB9xx.

Zhang, Y., Lee, T.S., Li, M., Liu, F., and Tang, S. (2019). Convolutional neural network models of V1 responses to complex patterns. J. Comput. Neurosci. *46*, 33–54.