Deep learning—Using machine learning to study biological vision

Najib J. Majaj
Center for Neural Science,
New York University, New York, NY, USA

Denis G. Pelli
Department of Psychology and Center for Neural Science,
New York University, New York, NY, USA

Many vision science studies employ machine learning, especially the version called "deep learning." Neuroscientists use machine learning to decode neural responses. Perception scientists try to understand how living organisms recognize objects. To them, deep neural networks offer benchmark accuracies for recognition of learned stimuli. Originally machine learning was inspired by the brain. Today, machine learning is used as a statistical tool to decode brain activity. Tomorrow, deep neural networks might become our best model of brain function. This brief overview of the use of machine learning in biological vision touches on its strengths, weaknesses, milestones, controversies, and current directions. Here, we hope to help vision scientists assess what role machine learning should play in their research.

Introduction

What does machine learning offer to biological vision scientists? It was developed as a tool for automated classification, optimized for accuracy. Machine learning is used in a broad range of applications (Brynjolfsson, 2018), from regression in stock market forecasting to reinforcement learning to play chess, but here we focus on classification. Physiologists use it to identify stimuli based on neural activity. To study perception, physiologists measure neural activity and psychophysicists measure overt responses, like pressing a button. Physiologists and psychophysicists are starting to consider deep learning as a model for object recognition by human and nonhuman primates (Cadieu et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014; Ziskind, Hénaff, LeCun, & Pelli, 2014; Testolin, Stoianov, & Zorzi, 2017). We suppose that most of our readers have heard of machine learning but are wondering whether it would be useful in their own research. We begin by describing some of its pluses and minuses.

Pluses: What it's good for

At the very least, machine learning is a powerful tool for interpreting biological data. A particular form of machine learning, *deep learning*, is very popular (Figure 1). Is it just a fad? For computer vision, the old paradigm was: feature detection, followed by segmentation, and then grouping (Marr, 1982). With machine learning tools, the new paradigm is to just define the task and provide a set of labeled examples, and the algorithm builds the classifier. (This is "supervised" learning; we discuss unsupervised learning below.)

Unlike the handcrafted pattern recognition (including segmentation and grouping) popular in the 70s and 80s, deep learning algorithms are generic, with little domain-specificity.[1] They replace hand-engineered feature detectors with filters that can be learned from the data. Advances in the mid-90s in machine learning made it useful for practical classification, such as handwriting recognition (LeCun et al., 1989; Vapnik, 2013).

Machine learning allows a neurophysiologist to decode neural activity without knowing the receptive fields (Seung & Sompolinsky, 1993; Hung. Kreiman, Poggio, & DiCarlo, 2005). Machine learning is a big step in the shifting emphasis in neuroscience from *how* the cells encode to *what* they encode—that is, what that code tells us about the stimulus (Barlow, 1953; Geisler, 1989). Mapping a receptive field is the foundation of neuroscience (beginning with Weber's 1834 mapping of tactile "sensory circles"). This once required single-cell recording, looking for minutes or hours at how one cell responds to each of perhaps a hundred different stimuli. Today it is clear that characterization

Citation: Majaj, N. J., & Pelli, D. G. (2018). Deep learning—Using machine learning to study biological vision. *Journal of Vision*, *18*(13):2, 1–13, https://doi.org/10.1167/18.13.2.

https://doi.org/10.1167/18.13.2 Received July 24, 2017; published December 3, 2018 ISSN 1534-7362

Copyright 2018 The Authors

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

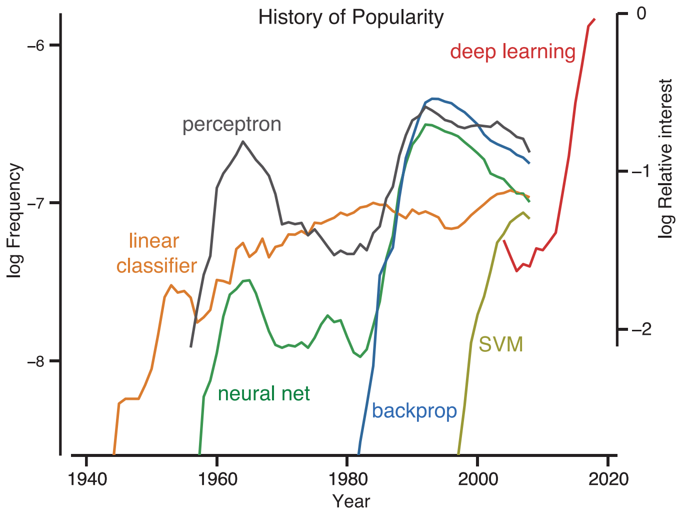

Figure 1. History of popularity. Lefthand scale: The frequency of appearance of each of five terms—linear classifier, perceptron, support vector machine, neural net, and backprop, (and *not* deep learning)—in books indexed by Google in each year of publication. Google counts instances of words and phrases of *n* words, and calls each an "ngram." Frequency is reported as a fraction of all instances of ngrams of that length, normalized by the number of books published that year (ngram / year / books published). The figure was created using Google's ngram viewer (https://books.google.com/ngrams), which contains a yearly count of ngrams found in sources printed between 1500 and 2008. Righthand scale: For deep learning, numbers represent worldwide search interest relative to the highest point on the chart for the given year for the term "deep learning" (as reported by https://trends.google.com/trends/). The righthand scale has been shifted vertically to match in 2004 the corresponding (not shown) deep learning ngram frequency (lefthand scale).

Glossary

Backprop: Short for "backward propagation of errors," it is widely used to apply gradient-descent learning to multilayer networks. It uses the chain rule from calculus to iteratively compute the gradient of the cost function for each layer.

Convexity: A real-valued function is called "convex" if the line segment between any two points on the graph of the function lies on or above the graph (Boyd & Vandenberghe, 2004). A problem is convex if its cost function is convex. Convexity guarantees that gradient descent will always find the global minimum.

Convolutional neural network (ConvNet): Rooted in the Neocognitron (Fukushima, 1980) and inspired by the simple and complex cells described by Hubel and Wiesel (1962), ConvNets apply backprop learning to multilayer neural networks based on convolution and pooling (LeCun et al., 1989; LeCun, Bottou, Bengio, & Haffner, 1998).

Cost function: A function that assigns a real number representing cost to a candidate solution by measuring the difference between the solution and the desired output. Solving by optimization means minimizing cost.

Cross-validation: Assesses the ability of the network to generalize from the data that it trained on to new data.

Deep learning: A successful and popular version of machine learning that uses backprop neural networks with multiple hidden layers. The 2012 success of AlexNet, then the best machine learning network for object recognition, was the tipping point. Deep learning is now ubiquitous in the Internet. The idea is to have each layer of processing perform successively more complex computations on the data to give the full multilayer network more expressive power. The drawback is that it is much harder to train multilayer networks (Goodfellow et al., 2016).

Generalization: How well a classifier performs on new, unseen examples that it did not see during training.

Gradient descent: An algorithm that minimizes cost by incrementally changing the parameters in the direction of steepest descent of the cost function.

Hebbian learning: According to Hebb's rule, the efficiency of a synapse increases after correlated pre- and post-synaptic activity. In other words, neurons that fire together, wire together (Lowel & Singer, 1992). Also known as spike-timing-dependent plasticity (Caporale & Dan, 2008).

Machine learning: Any computer algorithm that learns how to perform a task directly from examples, without a human providing explicit instructions or rules for how to do so. In one type of machine learning, called "supervised learning," correctly labeled examples are provided to the learning algorithm, which is then "trained" (i.e., its parameters are adjusted) to perform the task correctly on its own and generalize to unseen examples.

Neural nets: Computing systems inspired by biological neural networks that consist of individual neurons learning their connections with other neurons in order to solve tasks by considering examples.

Supervised learning: Any algorithm that accepts a set of labeled stimuli—a training set—and returns a classifier that can label stimuli similar to those in the training set.

Support vector machine (SVM): A type of machine learning algorithm for classification. An SVM uses the "kernel trick" to quickly learn to perform a nonlinear classification by finding a boundary in multidimensional space that separates different classes and maximizes the distance of class exemplars to the boundary (Cortes & Vapnik, 1995).

Unsupervised learning: Discovers structure and redundancy in data without labels. It is less widely used by computer scientists than supervised learning, but of great interest because labeled data are scarce while unlabeled data are plentiful.

of a single neuron's receptive field, which was invaluable in the retina and V1, fails to characterize how higher visual areas encode the stimulus. Machine learning techniques reveal "how neuronal responses can best be used (combined) to inform perceptual decision-making" (Graf, Kohn, Jazayeri, & Movshon, 2011). The simplicity of the machine decoding can be a virtue as it allows us to discover what can be easily read out (e.g., by a single downstream neuron; Hung et al. 2005). Achieving psychophysical levels of performance in decoding a stimulus object's identity and location from the neural response shows that the measured neural performance has all the information needed for the observer to do the task (Majaj, Hong, Solomon, & DiCarlo, 2015; Hong, Yamins, Majaj, & DiCarlo, 2016).

For psychophysics, signal detection theory (SDT) proved that the optimal classifier for a known signal in white noise is a template matcher (Peterson, Birdsall, & Fox, 1954; Tanner & Birdsall, 1958). Of course, SDT solves only a simple version of the general problem of object recognition. The simple version is for known signals, whereas the general problem includes variation in viewing conditions and diverse objects within a category (e.g., a chair can be any object that affords sitting). SDT introduces the very useful idea of a mathematically defined ideal observer, providing a reference for human performance (e.g., Geisler, 1989; Tjan & Legge, 1998; Pelli, Burns, Farell, & Moore-Page, 2006). However, one drawback is that it doesn't incorporate learning.

Deep learning, on the other hand, provides a pretty good observer that learns, which may inform studies of human learning.[2] In particular, it might reveal the constraints on learning imposed by the set of stimuli used in training. Further, unlike SDT, deep neural networks cope with the complexity of real tasks. It can be hard to tell whether behavioral performance is limited by the set of stimuli, their neural representation, or the observer's decision process (Majaj et al., 2015). Implications for classification performance are not readily apparent from direct inspection of families of stimuli and their neural responses. SDT specifies optimal performance for classification of known signals but does not tell us how to generalize beyond a training set. Machine learning does.

Minuses: Common complaints

Some biologists point out that neural nets do not match what we know about neurons (e.g., Crick, 1989; Rubinov, 2015; Heeger, 2017). Biological brains learn on the job, while neural networks need to converge before they can be used. Furthermore, once trained, deep networks generally compute in a feed-forward manner while there are major recurrent circuits in the cortex. But this may simply reflect the different ways that we use artificial and real neurons. The artificial networks are trained for a fixed task, whereas our visual brain must cope with a changing environment and task demands, so it never outgrows the need for the capacity to learn. Furthermore, there has recently been large progress in using trained recurrent neural networks both for computational tasks and as explanations for neural phenomena (Barak, 2017).

It is not clear, given what we know about neurons and neural plasticity, whether a backprop network can be implemented using biologically plausible circuits (but see Mazzoni, Andersen, & Jordan, 1991; Bengio, Le, Bornschein, Mesnard, & Lin, 2015). However, there are several promising efforts to implement more biological plausible learning rules, such as spike-timing–dependent plasticity (Mazzoni et al., 1991; Bengio et al., 2015; Sacramento, Costa, Bengio, & Senn, 2017).

Engineers and computer scientists, while inspired by biology, focus on developing machine learning tools that solve practical problems. Thus, models based on these tools often do not incorporate known constraints imposed by physiology. To this, one might counter that every biological model is an abstraction and can be useful even while failing to capture all the details of the living organism.

Some biological modelers complain that neural nets have alarmingly many parameters. Deep neural networks continue to be opaque. Before neural network modeling, a model was simpler than the data it explained. Deep neural nets are typically as complex as the data, and the solutions are hard to visualize (but see Zeiler & Fergus, 2013). While the training sets and learned weights are long lists, the generative rules for the network (the computer programs) are short. One flavor of this is proposals for cascaded canonical computations in the cortex (Hubel & Wiesel, 1962; Simoncelli & Heeger, 1998; Riesenhuber & Poggio, 1999; Serre, Wolf, Bileschi, Riesenhuber, & Poggio, 2007). Traditionally, having very many parameters has often led to overfitting—that is, good performance on the training set and poor performance beyond it—but the breakthrough is that deep-learning networks with a huge number of parameters nevertheless generalize well. Furthermore, Bayesian nonparametric models offer a disciplined approach to modeling with an unlimited number of parameters (Gershman & Blei, 2011). Mallat (2016) also notes that known symmetries of the problem can greatly reduce the number of parameters to be learned.

Some statisticians worry that rigorous statistical tools are being displaced by deep learning, which lacks rigor (Friedman, 1998; Matloff, 2014; but see Breiman, 2001; Efron & Hastie, 2016). Assumptions are rarely stated. There are no confidence intervals on the solution. However, performance is typically cross-validated, showing generalization. Deep learning is not convex, but it has been proven that convex networks can compute posterior probability (e.g., Rojas, 1996). Furthermore, machine learning and statistics seem to be converging to provide a more general perspective on rigorous probabilistic inference (Chung, Lee, & Sompolinsky, 2018).

Some physiologists note that decoding neural activity to recover the stimulus is interesting and useful but falls short of explaining what the neurons do. Some visual psychophysicists note some salient differences between performance of human observers and deep networks on tasks like object recognition and image distortion (Ullman, Assif, Fetava, & Harari, 2016; Berardino, Laparra, Ballé, & Simoncelli, 2017). Some cognitive psychologists dismiss deep neural networks as unable to "master some of the basic things that children do, like learning the past tense of a regular verb" (Marcus et al., 1992). Deep learning is slow. To recognize objects in natural images with the recognition accuracy of an adult, a state-of-the-art deep neural network needs 5,000 labeled examples per category (Goodfellow, Bengio, & Courville, 2016). But children and adults need only 100 labeled letters of an unfamiliar alphabet to reach the same accuracy as fluent native readers (Pelli et al., 2006). Overcoming these challenges may require more than deep learning.

These current limitations drive practitioners to enhance the scope and rigor of deep learning. But bear in mind that some of the best classifiers in computer science were inspired by biological principles (Rosenblatt, 1958; LeCun, 1985; Rumelhart, Hinton, & Williams, 1986; LeCun et al., 1989; Riesenhuber & Poggio, 1999; and see LeCun, Bengio, & Hinton 2015). Some of those classifiers are now so good that they occasionally exceed human performance and might serve as rough models for how biological systems classify (e.g., Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014; Ziskind et al., 2014; Testolin et al., 2017).

Milestones in classification

Mathematics versus engineering

The history of machine learning has two threads: mathematics and engineering (Figure 2). In the *mathematical* thread, two statisticians, Fisher (1922) and later Vapnik (2013), developed mathematical transformations to untangle categories. They assumed distributions and proved convergence.

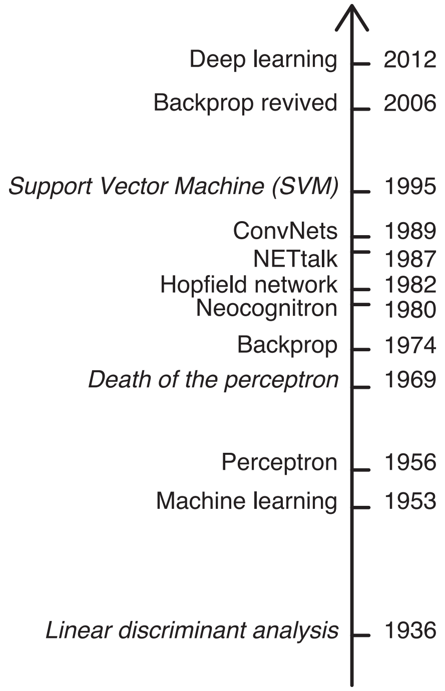

Figure 2. Milestones in classification. The math thread is italic; the engineering thread is plain.

In the *engineering* thread, a loose coalition of psychologists, neuroscientists, and computer scientists (e.g., Turing, Rosenblatt, Minsky, Fukushima, Hinton, Sejnowski, LeCun, Poggio, and Bengio) sought to reverse-engineer the brain to build a machine that learns. Their algorithms are typically applied to stimuli with unknown distributions and lack proofs of convergence.

1936: Linear discriminant analysis

Fisher (1936) introduced linear discriminant analysis to classify two species of iris flower based on four measurements per flower. When the distribution of the measurements is normal and the covariance matrix between the measurements is known, linear discriminant analysis answers the question: Supposing we use a single-valued function to classify, what linear function $y = w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4$, of four measurements x_1, x_2, x_3, and x_4 made on flowers, with free weights w_1, w_2, w_3, and w_4, will maximize discrimination of species?[3] Linear classifiers are great for simple problems for which the category boundary is a hyperplane in a small number of dimensions. However, complex problems like object recognition typically require more complex category boundaries in many dimensions. Furthermore, the distributions of the features are typically unknown and may not be normal.

Cortes and Vapnik (1995) noted that the first algorithm for pattern recognition was Fisher's optimal decision function for classifying vectors from two known distributions. Fisher solved for the optimal classifier in the presence of Gaussian noise and known covariance between elements of the vector. When the covariances are equal, this reduces to a linear classifier. The ideal template matcher of signal detection theory is an example of such a linear classifier (Peterson et al., 1954). This fully specified simple problem can be solved analytically. Of course, many important problems are not fully specified. In everyday perceptual tasks, we typically know only a "training" set of samples and labels.

1953: Machine learning

The first developments in machine learning were to play chess and checkers. "Could one make a machine to play chess, and to improve its play, game by game, profiting from its experience?" (Turing, 1953). Arthur Samuel (1959) defined *machine learning* as the "Field of study that gives computers the ability to learn without being explicitly programmed."

1958: Perceptron

Inspired by physiologically measured receptive fields, Rosenblatt (1958) showed that a very simple neural network, the perceptron, could learn to classify from training samples. Perceptrons combined several linear classifiers to implement piecewise-linear separating surfaces. The perceptron learns the weights to use in a linear combination of feature-detector outputs. The perceptron transforms the stimulus into a binary feature vector and then applies a linear classifier. The perceptron is piecewise linear and has the ability to learn from training examples without knowing the full distribution of the stimuli. Only the final layer in the perceptron learns.

1969: Death of the perceptron

However, it quickly became apparent that the perceptron and other single-layer neural networks cannot learn tasks that are not linearly separable—that is, they cannot solve problems like connectivity (Are all elements connected?) and parity (Is the number of elements odd or even?), which people solve readily (Minsky & Papert, 1988). On this basis, Minsky and Papert prematurely announced the death of artificial neural networks.

1974: Backprop

The death of the perceptron showed that learning in a one-layer network was too limited. This impasse was broken by the introduction of the backprop algorithm, which allowed learning to propagate through multiple-layer neural networks. The history of backprop is complicated (see Schmidhuber, 2015). The idea of minimization of error through a differentiable multistage network was discussed as early as the 1960s (e.g., Bryson, Denham, & Dreyfus, 1963). It was applied to artificial neural networks in the 1970s (e.g., Werbos, 1974). In the 1980s, efficient backprop first gained recognition, and led to a renaissance in the field of artificial neural network research (LeCun, 1985; Rumelhart, Hinton, & Williams, 1986). During the 2000s backprop neural networks fell out of favor, due to four limitations (Vapnik, 1999): (a) No proof of convergence. Backprop uses gradient descent. Gradient descent with a nonconvex cost function with multiple minima is only guaranteed to find a local, not the global minimum of the cost function. This has long been considered a major limitation, but LeCun et al. (2015) claim that it hardly matters in practice in current implementations of deep learning. (b) Slow. Convergence to a local minimum can be slow due to the high dimensionality of the weight space. (c) Poorly specified. Backprop neural networks had a reputation for being ill-specified, with an unconstrained number of units and training examples, and a step size that varied by problem. "Neural networks came to be painted as slow and fussy to train [,] beset by voodoo parameters and simply inferior to other approaches" (Cox & Dean, 2014). (d) Not biological. Lastly, backprop learning may not to be physiological: While there is ample evidence for Hebbian learning (increase of a synapse's gain in response to correlated activity of the two cells that it connects), such changes are never propagated backwards, beyond the one synapse, to a previous layer. With hindsight it is clear that a fifth limitation to the backprop in the 80s was inadequate resources: limited computing power and lack of large labeled datasets.

1980: Neocognitron, the first convolutional neural network

Fukushima (1980) proposed and implemented the Neocognitron, a hierarchical, multilayer artificial neural network. It recognized stimulus patterns (deformed numbers) despite small changes in position and shape.

1982: Hopfield network

The Hopfield network was introduced as a form of recurrent artificial network that serves as a contentaddressable memory and was proposed as a model for understanding human memory (Hopfield, 1982).

1987: NETtalk, the first impressive backprop neural network

Sejnowski and Rosenberg (1987) reported the exciting success of NETtalk, a neural network that learned to convert English text to speech:

> The performance of NETtalk has some similarities with observed human performance. (i) The learning follows a power law. (ii) The more words the network learns, the better it is at generalizing and correctly pronouncing new words. (iii) The performance of the networks degrades very slowly as connections in the network are damaged: no single link or processing unit is essential. (iv) Relearning after damage is much faster than learning during the original training…

1989: ConvNets

Yann LeCun and his colleagues combined convolutional neural networks with backprop to recognize handwritten characters (LeCun et al., 1989). This network was commercially deployed by AT&T, and today reads millions of checks a day (LeCun et al., 1998). Later, adding half-wave rectification and max pooling greatly improved its accuracy in recognizing objects (Jarrett, Kavukcuoglu, & LeCun, 2009).

1995: Support vector machine (SVM)

Cortes & Vapnik (1995) proposed the support vector network, a learning machine for binary classification problems. Support vector machines (SVMs) generalize well and are free of mysterious training parameters. Some versions of the SVM are convex (e.g., Lin, 2001).

2006: Backprop revived

Hinton and Salakhutdinov (2006) sped up backprop learning by unsupervised pretraining. This helped to revive interest in backprop (Hinton, Osindero, & Teh, 2006). In the same year, a supervised backprop-trained convolutional neural network set a new record on the famous MNIST handwritten-digit recognition benchmark (Ranzato et al., 2006, 2007).

2012: Deep learning

Geoff Hinton said, “It took 17 years to get deep learning right; one year thinking and 16 years of progress in computing, praise be to Intel” (Cox & Dean, 2014; LeCun et al., 2015). It is not clear who coined the term “deep learning.”[4] In their book, *Deep Learning Methods and Applications*, Deng and Yu (2014) cite Hinton et al. (2006) and Bengio (2009) as the first to use the term. However, the big debut for deep learning was an influential paper by Krizhevsky, Sutskever, and Hinton (2012) describing AlexNet, a deep convolutional neural network that classified 1.2 million high-resolution images into 1,000 different classes, greatly outperforming previous state-of-the-art machine learning and classification algorithms.

Controversies

The field is growing quickly, yet certain topics remain hot. For proponents of deep learning, the ideal network is composed of simple elements and learns everything from the training data. At the other extreme, computer vision scientists argue that we know a lot about how the brain recognizes objects, which we can engineer into the networks before learning (e.g., gain control and normalization). Some engineers look to the brain only to copy strengths of the biological solution; some scientists think there are useful clues in its limitations as well (e.g., crowding).

Is deep learning the best solution for all visual tasks?

Deep learning is not the only thing in the vision scientist’s toolbox. The complexity of deep learning may be unwarranted for simple problems that are well handled by, for example, SVM. Try shallow networks first, and if they fail, go deep.

Why object recognition?

The visual task of object recognition has been very useful in vision research because it is an objective task that is easily scored as right or wrong, is essential in daily life, and captures some of the magic of seeing. It is a classic problem with a rich literature. Deep neural nets solve it, albeit with a million parameters. Recognizing objects is a basic life skill, including recognition of words, people, things, and emotions. The concern that the research focus on object recognition might be merely an obsession of the scientists rather than a central task of biological vision is countered by hints that visual perception is biased to interpret the world as consisting of discrete objects even when it isn't, such as when we see animals in the clouds.

Of course, there are many other important visual tasks, including interpolation (e.g., filling in) and extrapolation (e.g., estimating heading). The inverse of categorization is synthesis. Human estimation of one feature, such as brightness or speed, is imprecise and adequately represented by roughly seven categories (Miller, 1956). For detection of image distortion, a simple model with gain-control normalization is better than current deep networks (Berardino et al., 2017). Scientists, like the brain, use whatever tool works best.

Deep learning is not convex

A problem is convex if its cost function is convex—that is, if the line between any two points on the function lies on or above the function. This guarantees that gradient descent will find the global minimum. For some combinations of stimuli, categories, and classifiers, convexity can be proven. In machine learning, some kernel methods, including SVMs, have the advantage of convexity, at the cost of limited generalization. In the 1990s, SVMs were popular because they guaranteed fast convergence even with a large number of training samples (Cortes & Vapnik, 1995). However, cost functions for deep neural networks are not convex. Unlike convex functions, nonconvex functions can have multiple minima and saddle points. The challenge in high dimensional cost functions is the saddle points, which greatly outnumber the local minima, but there are tricks for not getting stuck at saddle points (Dauphin et al., 2014). Although deep neural networks are not convex, they do fit the training data and generalize well (LeCun, Bengio, & Hinton, 2015).

Shallow versus deep networks

The field's imagination has focused alternately on shallow and deep networks, beginning with the perceptron in which only one layer learned, followed by backprop, which allowed multiple layers to learn, and cleared the hurdles that doomed the perceptron. Then SVM, with its single layer, sidelined the multilayer backprop. Today multilayer deep learning reigns; Krizhevsky et al. (2012) attributed the success of AlexNet to its eight-layer depth; it performed worse with fewer layers. Some people claim that deep learning is essential to recognize objects in real-world scenes. For example, the "Inception" 22-layer deep learning network won the Image Net Real World Challenge in 2014 (Szegedy et al., 2015).

The need for depth is hard to prove, but, in considering the depth versus width of a feed-forward neural network, Eldan and Shamir (2016) showed that a radial function can be approximated by a three-layer network with far fewer neurons than the best two-layer network (also see Telgarsky, 2015). Object recognition implies a classification function that assigns one of several discrete values to each image. Mhaskar, Liao, and Poggio (2017) suggest that for real-world recognition the classification function is typically compositional—that is, a hierarchy of functions, one per node, in feed-forward layers, in which the receptive fields of higher layers are ever larger. They argued that scalability and shift invariance in natural images require compositional algorithms. They prove that deep hierarchical networks can approximate compositional functions with the same accuracy as shallow networks but with exponentially fewer training parameters.

Supervised versus unsupervised

Learning algorithms for a classifier can be supervised or not (i.e., need labels for training, or don't). Today most machine learning is supervised (LeCun, Bengio, & Hinton, 2015). The images are labeled (e.g., "car" or "face"), or the network receives feedback on each trial from a cost function that assesses how well its answer matches the image's category. In unsupervised learning, no labels are given. The algorithm processes images, typically to minimize error in reconstruction, with no extra information about what is in the (unlabeled) image. A cost function can also reward decorrelation and sparseness (e.g., Olshausen & Field, 1996). This allows learning of image statistics and has been used to train early layers in deep neural networks. Human learning of categorization is sometimes done with explicitly named objects—"Look at the tree!"—but more commonly the feedback is implicit. Consider reaching your hand to raise a glass of water to your lips. Contact informs vision. On specific benchmarks, where the task is well-defined and labeled examples are available, supervised learning can excel (e.g., AlexNet), but unsupervised learning may be more useful when few labels are available. Unsupervised learning adjusts the network to suit the statistics of the world (Hinton & Salakhutdinov, 2006).

Current directions

What does deep learning add to the vision science toolbox?

Deep learning is more than just a souped-up regression (Marblestone, Wayne, & Kording, 2016).

Like SDT, it allows us to see more in our behavioral and neural data. In the 1940s, Norbert Wiener and others developed algorithms to automate and optimize signal detection and classification. A lot of it was engineering. The whole picture changed with the SDT theorems, mainly the proof that the maximum-likelihood receiver is optimal for a wide range of simple tasks (Peterson et al., 1954). In white noise a traditional receptive field computes the likelihood of the presence of a signal matching the receptive field weights. It was exciting to realize that the brain contains 10^{11} likelihood computers. Later work added prior probability, for a Bayesian approach. Tanner and Birdsall (1958) noted that, when figuring out how a biological system does a task, it is very helpful to know the optimal algorithm and to rate observed performance by its *efficiency* relative to the optimum. SDT solved detection and classification mathematically, as maximum likelihood. It was the classification math of the 60s. Machine learning is the classification math of today. Both enable deeper insight into how biological systems classify. Of course, as noted above, SDT is restricted to the case of known signals in additive noise, whereas deep learning can solve real-world object recognition like detecting a dog in a snapshot after training on labeled examples. In the old days we used to compare human and ideal classification performance (Pelli et al., 2006). Today, we also compare human and machine learning. Deep learning is the best model we have today for how complex systems of simple units can recognize objects as well as the brain does. Several labs are currently comparing patterns of activity of particular artificial layers to neural responses in various cortical areas of the mammalian visual brain (Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al. 2014; Schrimpf et al. 2018).

What can computer scientists learn from psychophysics?

Computer scientists build classifiers to recognize objects. Vision scientists, including psychologists and neuroscientists, study how people and animals classify in order to understand how the brain works. So, what do computer and vision scientists have to say to each other? Machine learning accepts a set of labeled stimuli to produce a classifier. Much progress has been made in physiology and psychophysics by characterizing how well biological systems can classify stimuli. The psychophysical tools (e.g., threshold and SDT) developed to characterize behavioral classification performance are immediately applicable to characterize classifiers produced by machine learning (e.g., Ziskind, Hénaff, LeCun, & Pelli, 2014; Testolin, Stoianov, & Zorzi, 2017).

Psychophysics

"Adversarial" examples have been presented as a major flaw in deep neural networks (Hutson, 2018; Mims, 2018). These slightly doctored images of objects are misclassified by a trained network, even though the doctoring has little effect on human observers. The same doctored images are similarly misclassified by several different networks trained with the same stimuli (Szegedy et al., 2013). Humans too have adversarial examples. Illusions are robust classification errors (Shapiro & Todorovic, 2016). The blindspot-filling-in illusion is a dramatic adversarial example in human vision. While viewing with one eye, two fingertips touching in the blindspot are perceived as one long finger. If the image is shifted a bit so that the fingertips emerge from the blindspot, the viewer sees two fingers. Neural networks lacking the anatomical blindspot of human vision are hardly affected by the shift (but see Azulay & Weiss, 2018). The existence of adversarial examples is intrinsic to classifiers trained with finite data, whether biological or not. In the absence of information, neural networks interpolate and so do biological brains. Psychophysics, the scientific study of perception, has achieved its greatest advances by studying classification errors (Fechner, 1860). Such errors can reveal blindspots. Stimuli that are physically different yet indistinguishable are called *metamers*. The systematic understanding of color metamers revealed the three dimensions of human color vision (Palmer, 1777; Young, 1802; Helmholtz, 1867). In recent work, many classifiers have been trained solely with the objects they are meant to classify, and thus will classify everything as one of those categories, even doctored noise that is very different from all of the images. It is important to train with sample images that represent the entire test set.

Conclusion

Machine learning is here to stay. Deep learning is better than the "neural" networks of the 80s. Machine learning is useful both as a model for perceptual processing, and as a decoder of neural processing, to see what information the neurons are carrying. The large size of the human cortex is a distinctive feature of our species and crucial for learning. It is anatomically homogenous yet solves diverse sensory, motor, and cognitive problems. Key biological details of cortical learning remain obscure, but, even if they ultimately preclude backprop, the performance of current machine learning algorithms is a useful benchmark.

Resources

We recommend textbooks on deep learning by Goodfellow, Bengio, and Courville (2016) and Ng (2017). There are many packages for optimization and machine learning in MATLAB and Python.

Keywords: deep learning, machine learning, neural networks, object recognition

Acknowledgments

Thanks to Yann LeCun for helpful conversations. Thanks to Aenne Brielmann, Augustin Burchell, Kaitie Holman, Katerina Malakhova, Cristina Savin, Laura Suciu, Pascal Wallisch, and Avi Ziskind for helpful comments on the manuscript. NJM was supported by National Institute of Health (NIH) R01 grants EY005846 and EY024914. DGP was supported by NIH grant R01 EY027964.

Commercial relationships: none.
Corresponding author: Najib J. Majaj.
E-mail: najib.majaj@nyu.edu.
Address: Department of Psychology and Center for Neural Science, New York University, New York, NY, USA.

Footnotes

[1] Admittedly, these networks still demand tweaking of a few parameters, including number of layers and number of units per layer.

[2] In the same spirit, "sequential ideal observer" and "accuracy maximization" model generalized ideal observer calculations to include a shallow form of supervised learning (Geisler, 1989; Burge & Jaini, 2017).

[3] Linear discriminant analysis is an outgrowth of regression, which has a much longer history. Regression is the optimal least-squares linear combination of given functions to fit given data and was applied by Legendre (1805) and Gauss (1809) to astronomical data to determine the orbits of the comets and planets around the sun. The estimates come with confidence intervals and the fraction of variance accounted for, which rates the goodness of the explanation.

[4] The idea of deep learning is not exclusive to machine learning and neural networks (e.g., Dechter, 1986).

References

Azulay, A., & Weiss, Y. (2018). Why do deep convolutional networks generalize so poorly to small image transformations? *arXiv*:1805.12177.

Barak, O. (2017). Recurrent neural networks as versatile tools of neuroscience research. *Current Opinion in Neurobiology*, *46*, 1–6.

Barlow, H. B. (1953). Summation and inhibition in the frog's retina. *The Journal of Physiology*, *119*(1), 69–88.

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, *2*(1), 1–127.

Bengio, Y., Le, D. H., Bornschein, J., Mesnard, T., & Lin, Z. (2015). Towards biologically plausible deep learning. *arXiv*:1502.04156.

Berardino, A., Laparra, V., Ballé, J., & Simoncelli, E. (2017). Eigen-distortions of hierarchical representations. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30* (pp. 3533–3542). Red Hook, NY: Curran Associates, Inc. https://papers.nips.cc/book/advances-in-neural-information-processing-systems-30-2017

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge, UK: Cambridge University Press.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231.

Brynjolfsson, E. (2018, July 26). Machine learning will be the engine of global growth. *Financial Times*. https://www.ft.com/content/133dc9c8-90ac-11e8-9609-3d3b945e78cf

Bryson, A. E., Denham, W. F., & Dreyfus, S. E. (1963). Optimal programming problems with inequality constraints. *AIAA Journal*, *1*(11), 2544–2550.

Burge, J., & Jaini, P. (2017). Accuracy maximization analysis for sensory-perceptual tasks: Computational improvements, filter robustness, and coding advantages for scaled additive noise. *PLoS Computational Biology*, *13*(2), e1005281.

Cadieu, C. F., Hong, H., Yamins, D. L., 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. *PLoS Computational Biology*, *10*(12), e1003963.

Caporale, N., & Dan, Y. (2008). Spike timing-dependent plasticity: A Hebbian learning rule. *Annual Review of Neuroscience*, *31*, 25–46.

Chung, S., Lee, D. D., & Sompolinsky, H. (2018). Classification and geometry of general perceptual manifolds. *Physical Review X*, *8*(3), 031003.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Cox, D. D., & Dean, T. (2014). Neural networks and neuroscience-inspired computer vision. *Current Biology*, *24*(18), R921–R929.

Crick, F. (1989, January 12). The recent excitement about neural networks. *Nature*, *337*(6203), 129–132.

Dauphin, Y. N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., & Bengio, Y. (2014). Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 27* (pp. 2933–2941). Red Hook, NY: Curran Associates, Inc. Retrieved from https://papers.nips.cc/book/advances-in-neural-information-processing-systems-27-2014

Dechter, R. (1986). Learning while searching in constraint-satisfaction-problems. In *Proceedings of the Fifth AAAI National Conference on Artificial Intelligence* (pp. 178–183). AAAI Press.

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

Efron, B., & Hastie, T. (2016). *Computer age statistical inference: Algorithms, evidence, and data science* (Vol. 5). Cambridge University Press.

Eldan, R., & Shamir, O. (2016). The power of depth for feedforward neural networks. In *Conference on learning theory* (pp. 907–940).

Fechner, G. T. (1860). *Elemente der psychophysik*. Leipzig, Germany: Breitkopf and Hartel. Translated by H. E. Adler, D. H. Howes, & E. G. Boring (1966). *Elements of psychophysics*. New York, NY: Holt, Rinehart and Winston.

Fisher, R. A. (1922). The goodness of fit of regression formulae, and the distribution of regression coefficients. *Journal of the Royal Statistical Society*, *85*(4), 597–612, https://doi.org/10.2307/2341124.

Friedman, J. H. (1998). Data mining and statistics: What's the connection? *Computing Science and Statistics*, *29*(1), 3–9.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 193–202.

Gauss, C.F. (1809). *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*. Translated by C. H. Davis (1857): *Theory of the motion of the heavenly bodies moving about the sun in conic sections*. Boston, MA: Little, Brown and Company.

Geisler, W. S. (1989). Sequential ideal-observer analysis of visual discriminations. *Psychological Review*, *96*(2), 267.

Gershman, S. J., & Blei, D. M., (2011). A tutorial on Bayesian nonparametric models. *arXiv*:1106.2697 [stat.ML].

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge: MIT Press.

Graf, A. B., Kohn, M., Jazayeri, J., & Movshon, J. A. (2011). Decoding the activity of neuronal populations in macaque primary visual cortex. *Nature Neuroscience*, *14*(2), 239–247.

Heeger, D. J. (2017). Theory of cortical function. *Proceedings of the National Academy of Sciences*, *114*(8), 1773–1782.

Helmholtz, H. (1867). *Handbuch der physiologischen Optik* [*Helmholtz's treatise on physiological optics*] (Trans. 1924 from the 3rd German ed.; James P. C. Southall, Ed.). Rochester, NY: Optical Society of America.

Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*(7), 1527–1554.

Hinton, G. E., & Salakhutdinov, R. R. (2006, July 28). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507.

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.

Hopfield, J. J. (1982) Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*(8), 2554–2558.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*(1), 106–154.

Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005, November 4). Fast readout of object identity from macaque inferior temporal cortex. *Science*, *310*(5749), 863–866.

Hutson, M. (2018, July 20). Hackers easily fool artificial intelligences. *Science*, *361*(6399), 215.

Jarrett, K., Kavukcuoglu, K., & LeCun, Y. (2009). What is the best multi-stage architecture for object recognition? In *IEEE 12th International Conference on Computer Vision* (pp. 2146–2153). https://ieeexplore.ieee.org/abstract/document/5459469

Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014).

Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, *10*(11), e1003915.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, (pp. 1097–1105). Red Hook, NY: Curran Associates, Inc. Retrieved from https://papers.nips.cc/book/advances-in-neural-information-processing-systems-25-2012

LeCun, Y. (1985). Une procedure d'apprentissage pour reseau a seuil asymmetrique [A learning scheme for asymmetric threshold networks]. In *Proceedings of Cognitiva* (pp. 599–604). Paris, France.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*(4), 541–551.

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

Legendre, A.-M. (1805). *Nouvelles méthodes pour la détermination des orbites des comètes* [New methods for the determination of the orbits of comets]. Paris, France: F. Didot.

Lin, C. J. (2001). On the convergence of the decomposition method for support vector machines. *IEEE Transactions on Neural Networks*, *12*(6), 1288–1298.

Lowel, S., & Singer, W. (1992, January 10). Selection of intrinsic horizontal connections in the visual cortex by correlated neuronal activity. *Science*, *255*(5041), 209.

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.

Mallat, S. (2016). Understanding deep convolutional networks. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *374*(2065):20150203.

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., Xu, F., & Clahsen, H. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, *57*(4), I-178, https://doi.org/10.2307/1166115.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco, CA: Freeman and Company.

Matloff, N. (2014, August 26). Statistics: Losing ground to CS, losing image among students. *Revolutions*. Retrieved from http://blog.revolutionanalytics.com/2014/08/statistics-losing-ground-to-cs-losing-image-among-students.html

Marblestone, A. H., Wayne, G., & Kording, K. P. (2016). Toward an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience*, *10*.

Mazzoni, P., Andersen, R. A., & Jordan, M. I. (1991). A more biologically plausible learning rule for neural networks. *Proceedings of the National Academy of Sciences*, *88*(10), 4433–4437.

Mhaskar, H., Liao, Q., & Poggio, T. A. (2017). When and why are deep networks better than shallow ones? In *AAAI* (pp. 2343–2349).

Miller, G. A. (1956). The magic number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 91–97.

Mims, C. (2018, August 4). Should artificial intelligence copy the human brain? *The Wall Street Journal*. Retrieved from https://www.wsj.com/articles/should-artificial-intelligence-copy-the-human-brain-1533355265

Minsky, M., & Papert, S. (1988). *Perceptrons: An introduction to computational geometry*. Cambridge, MA: MIT Press.

Ng, A. (2017). *Machine learning yearning*. Retrieved from http://www.mlyearning.org/

Olshausen, B. A., & Field, D. J. (1996, June 13). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), 607.

Palmer, G. (1777). *Theory of colour and vision*. London: Leacroft.

Pelli, D. G., Burns, C. W., Farell, B., & Moore-Page, D. C. (2006). Feature detection and letter identification. *Vision Research*, *46*(28), 4646–4674.

Peterson, W. W. T. G., Birdsall, T., & Fox, W. (1954). The theory of signal detectability. *Transactions of the IRE Professional Group on Information Theory*, *4*(4), 171–212.

Ranzato, M. A., Huang, F. J., Boureau, Y. L., & LeCun, Y. (2007). Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *2007 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–8).

Ranzato, M. A., Poultney, C., Chopra, S., & LeCun, Y. (2006). Efficient learning of sparse representations with an energy-based model. In J. C. Schölkopf, J. C. Platt, T. Hoffman, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 19*, (pp. 1137–1144). Red Hook, NY: Curran Associates, Inc. Retrieved from https://papers.nips.cc/book/advances-in-neural-information-processing-systems-19-2006.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature neuroscience*, *2*(11), 1019–1025.

Rojas, R. (1996). A short proof of the posterior probability property of classifier neural networks. *Neural Computation*, *8*(1), 41–43.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408.

Rubinov, M. (2015). Neural networks in the future of neuroscience research. *Nature Reviews Neuroscience*, *16*(12), 767.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533–536, https://doi.org/10.1038/323533a0.

Sacramento, J., Costa, R. P., Bengio, Y., & Senn, W. (2017). Dendritic error backpropagation in deep cortical microcircuits. *arXiv*:1801.00062.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *3*(3), 210–229.

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

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N., Rajalingham, R., Issa, . . . DiCarlo, J. J. (2018). GBrain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? *bioRxiv*:407007.

Sejnowski, T. J., & Rosenberg, C. R. (1987). Parallel networks that learn to pronounce English text. *Complex Systems*, *1*(1), 145–168.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *3*, 411–426.

Seung, H. S., & Sompolinsky, H. (1993). Simple models for reading neuronal population codes. *Proceedings of the National Academy of Sciences*, *90*(22), 10749–10753.

Shapiro, A. G., & Todorovic, D. (Eds.). (2016). *The Oxford compendium of visual illusions*. New York, NY: Oxford University Press.

Simoncelli, E. P., & Heeger, D. J. (1998). A model of neuronal responses in visual area MT. *Vision Research*, *38*(5), 743–761.

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

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

Tanner W. P., Jr., & Birdsall, T. G. (1958). Definitions of d' and η as psychophysical measures. *The Journal of the Acoustical society of America*, *30*(10), 922–928.

Telgarsky, M. (2015). Representation benefits of deep feedforward networks. *arXiv*:1509.08101.

Testolin, A., Stoianov, I., & Zorzi, M. (2017). Letter perception emerges from unsupervised deep learning and recycling of natural image features. *Nature Human Behaviour*, *1*(9), 657.

Tjan, B. S., & Legge, G. E. (1998). The viewpoint complexity of an object-recognition task. *Vision Research*, *38*(15–16), 2335–2350.

Turing, A. M. (1953). Digital computers applied to games. In B. V. Bowden (Ed.), *Faster than thought*. London: Pitman.

Ullman, S., Assif, L., Fetaya, E., & Harari, D. (2016). Atoms of recognition in human and computer vision. *Proceedings of the National Academy of Sciences*, *113*(10), 2744–2749.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE transactions on neural networks*, *10*(5), 988–999.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer Science & Business Media.

Weber, E. H. (1834). De pulsu, resorptione, auditu et tactu: Annotationes anatomicae et physiologicae. Printed by Koehler. Leipzig, Germany. Edited and Translated by H. E. Ross & D. J. Murray (1996) as E.H. Weber: On the tactile senses, Erlbaum (UK) Taylor Francis, Publishers. East Sussex, UK.

Werbos, P. (1974). *Beyond regression: New tools for prediction and analysis in the behavioral sciences* (Unpublished doctoral dissertation). Harvard University, Cambridge, MA.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.

Young, T. (1802). The Bakerian lecture: On the theory of light and colours. *Philosophical Transactions of* *the Royal Society of London*, *92*, 12–48, https://doi.org/10.1098/rstl.1802.0004.

Yule, G. U. (1897). On the theory of correlation. *Journal of the Royal Statistical Society*, *60*(4), 812–854.

Zeiler, M. D., & Fergus, R. (2013). Visualizing and understanding convolutional networks. *arXiv*: 1311.2901.

Ziskind, A.J., Hénaff, O., LeCun, Y., & Pelli, D.G. (2014) *The bottleneck in human letter recognition: A computational model*. Poster presented at the Vision Sciences Society, St. Pete Beach, Florida. Retrieved from http://f1000.com/posters/browse/summary/1095738