Ann. N.Y. Acad. Sci. ISSN 0077-8923

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES
Special Issue: *The Year in Cognitive Neuroscience*

Review

From convolutional neural networks to models of higher-level cognition (and back again)

Ruairidh M. Battleday,[1,a] Joshua C. Peterson,[1,a] and Thomas L. Griffiths[1,2]

[1]Department of Computer Science, Princeton University, Princeton, New Jersey. [2]Department of Psychology, Princeton University, Princeton, New Jersey

Addresses for correspondence: Ruairidh M. Battleday and Joshua C. Peterson, Department of Computer Science, Princeton University, 35 Olden Street, Princeton, NJ 08540. battleday@princeton.edu and joshuacp@princeton.edu

The remarkable successes of convolutional neural networks (CNNs) in modern computer vision are by now well known, and they are increasingly being explored as computational models of the human visual system. In this paper, we ask whether CNNs might also provide a basis for modeling higher-level cognition, focusing on the core phenomena of similarity and categorization. The most important advance comes from the ability of CNNs to learn high-dimensional representations of complex naturalistic images, substantially extending the scope of traditional cognitive models that were previously only evaluated with simple artificial stimuli. In all cases, the most successful combinations arise when CNN representations are used with cognitive models that have the capacity to transform them to better fit human behavior. One consequence of these insights is a toolkit for the integration of cognitively motivated constraints back into CNN training paradigms in computer vision and machine learning, and we review cases where this leads to improved performance. A second consequence is a roadmap for how CNNs and cognitive models can be more fully integrated in the future, allowing for flexible end-to-end algorithms that can learn representations from data while still retaining the structured behavior characteristic of human cognition.

Keywords: categorization; cognitive modeling; convolutional neural networks; similarity; vision

Introduction

The first demonstration of the potential impact of deep learning came from the field of computer vision, with the unprecedented success of a deep convolutional neural network (CNN) called "AlexNet" on the ImageNet challenge—a large-scale natural image classification task.[1] Through a number of scientific and engineering advances, researchers were able to train much larger artificial neural networks than previously, and in doing so apply them to a broader and more difficult range of tasks. Beyond making networks deeper, success was found to depend on matching the right types of architectural and functional constraints to different kinds of task—collectively known as *inductive biases*, or the aspects of model design that influence the generalization performance of a learning algorithm beyond the data themselves.[2] For computer vision tasks involving images, the inductive bias provided by deep stacks of learnable convolutional filters has proved instrumental in reaching state-of-the-art performance on a large number of computer vision benchmarks related to cognition, such as image classification,[1,3,4] image segmentation and object recognition,[5,6] and visual question answering;[7–10] as well as in visual modules for algorithms solving more integrated tasks.[11] Developments in CNN structure and applications continue at an ever increasing rate in the computer vision and machine learning communities, to the point where it has become rare to hear of a statistical tool or application related to

[a]Both of these authors contributed equally.

doi: 10.1111/nyas.14593

image processing that does not include a CNN component.

In addition to solving problems in computer vision, there has been increasing recognition that CNNs could also serve as models of the human visual system.[12] CNNs have repeatedly proved to be the best predictive model of neural and voxel responses to visual stimuli in primate electrophysiology and human imaging studies, with the degree of approximation roughly improving with CNN performance.[13–22] Interestingly, the level of cortical and CNN hierarchy often appears to match, with shallow CNN layers better predicting earlier visual areas and deeper CNN layers better predicting higher visual areas.[15,17,18]

CNNs represent an equally interesting opportunity for the study of higher-level cognition. That is, they may allow us to extend computational models of cognition from simple artificial stimuli to more complex naturalistic stimuli representative of the complex visual world in which our cognitive abilities arise. In particular, cognitive psychology has provided a range of influential computational models that examine the fundamental cognitive phenomena of similarity and categorization, which typically abstract over details of neural implementation, and instead focus on explaining human behavior in terms of the computational problems posed by the environment and cognitive processes that solve it.[23,24] However, well-motivated considerations—namely, having good representations of stimuli on which to base quantitative analysis—have limited these models to examining behavior from laboratory-based experimental paradigms using simple artificial stimuli. The rich representations produced by CNNs solving related computational tasks suggest they are a good candidate to extend this modeling framework further, to more naturalistic stimuli. If effective, this would allow us to study human behavior in more ecologically relevant settings—a goal that must be an ultimate aim of any research program into the mind and brain.

In this review, we examine how CNNs can be used to extend the range of traditional models of higher-level cognition, and, in turn, whether insights from studies of cognition can improve the performance of CNNs in solving related computer vision and machine learning tasks (Fig. 1). We begin with an introduction to CNNs, and then show how representations from these networks can be directly integrated into existing cognitive models of similarity and categorization as tools for feature learning over diverse ranges of naturalistic stimuli. Although off-the-shelf CNN representations prove our best available substrate for modeling behavior, a growing number of cognitive models have been proposed that increase their correspondence to psychological representations using simple mathematical transformations. Beyond capturing human behavior better, these modeling regimes give us tools to integrate more structured, cognitive constraints into normal CNN training paradigms, and we review a number of examples where this has led to improved performance. Finally, we give an account of how deep learning and higher-level cognitive modeling can be more fully integrated in the future, in order to obtain the advantages of both frameworks and provide the next generation of cognitive models.

Deep and convolutional neural networks

Artificial neural networks have a long connection to cognitive neuroscience, both as models of neurons and as a brain-compatible computing paradigm. Interest began with the investigation of the computational properties of simplified models of single neurons in the mid-20th century,[25,26] inspired by theories about the biology of neurons.[27,28] One goal was to emulate the cognitive properties of the brain that had remained elusive for symbolic approaches: good performance on inductive learning tasks, robustness to error in the input or computation, and graceful degradation of performance with alteration of the type of input or task. A second wave of research was driven by learning algorithms that allowed researchers to train the hidden weights in multilayer networks automatically, including the "backpropagation" algorithm.[29] Learning these hidden weights allowed networks to solve nonlinear problems, with research moving quickly from demonstrations on "toy" logical problems like implementing the XOR function to more cognitively interesting tasks. Key to the important series of successes in this era was the identification of different inductive biases as instrumental to solving different kinds of problems—for example, the need for recurrent loops in time-dependent tasks,[30] certain intermediate forms in linguistics,[31] and convolutional filters in image classification problems.[32]

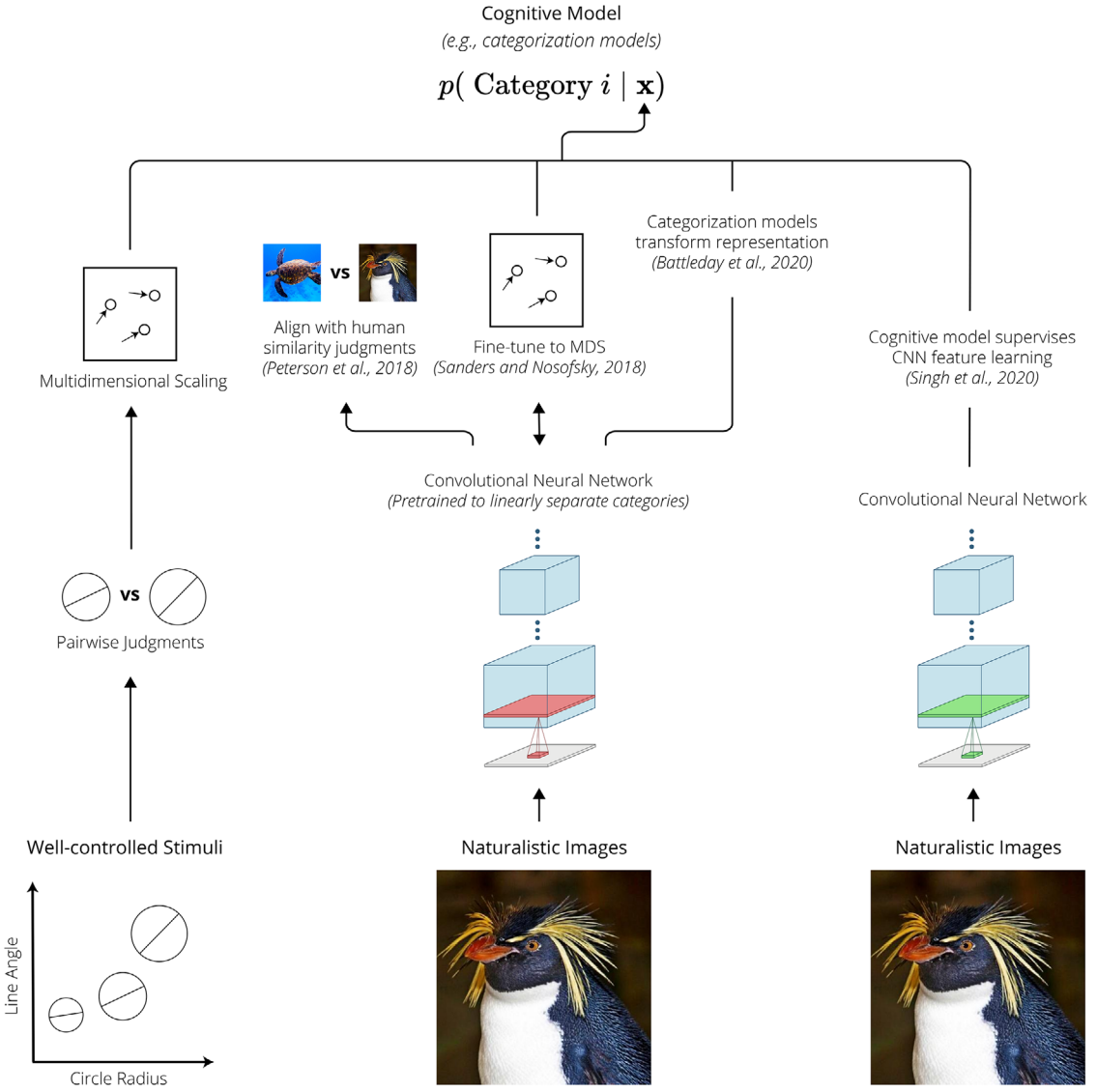

Figure 1. **Overview. Traditional studies have used simple artificial stimuli that can be mathematically represented unambiguously as the substrate for models of higher-level cognition (left pathway). CNNs can be used to supply representations for more complex naturalistic images, which can be further modified to better reflect human judgments before being input into the same kinds of cognitive model (middle pathways).[50,64,73] End-to-end models offer the opportunity to solve both of these problems simultaneously and learn a representation for naturalistic stimuli that satisfies the constraints inherent in higher-level cognitive models (right pathway).[136]**

Many other components of the modern deep learning framework arose during this period from the collaboration of psychologists, neuroscientists, and the computational vision community, including the development of hierarchical feed-forward visual models based on stacks of nonlinear feature maps and pooling between layers.[33–36]

The third wave of research into artificial neural networks, now known as the "deep learning revolution," was heralded in 2012 by the unprecedented performance of a CNN with several hidden layers, "AlexNet," on a difficult task of natural image classification.[1] Not only did the large improvement in classification accuracy between AlexNet

and previous state-of-the-art classifiers signal interest, but also the rapid subsequent improvements in performance seemed to announce that human-level performance on human-relevant tasks might be in sight. Although most of the components of AlexNet came from the first two waves of interest in artificial neural networks, it took a number of engineering and scientific advances to render performance at this scale achievable. First, these successes came hand-in-hand with exponentially increased computational power and the release of very large datasets (e.g., ImageNet).[1,3] Having more data and the computational power to approximate more complex functions meant the range of tasks that deep neural networks could learn to perform was significantly extended. But this belies the second, major influence, which is the scientific and societal side of the revolution. Initial successes drew the attention of a large community versed in statistics and computation, and the sheer breadth of architectures, learning rules, objective functions, and datasets tested by the community became an important factor in itself. Focus is now leveled not at laboratory problems, but at real-world problem-solving environments, such as the prediction and generation of text,[37] performance on video games,[11] and large-scale categorization of natural images.[1,3]

The deep neural networks that have provided the closest correspondence with brain and behavioral data using visual stimuli have been trained in the context of classification by supervised learning,[12,38] which refers to training a learning algorithm to improve its predictions by comparing them to the known correct answer—a form of teaching, or "supervision." This comparison is given explicit form using an objective function, and the algorithm must use the error signal it returns to maximize its performance. In classification tasks, the input is usually some naturalistic stimulus, such as an image or text, encoded as a matrix or vector of activation values, and the output most commonly a category label. Deep neural networks solve this problem by successively transforming activations using a sequential stack of feed-forward layers, ending with a category label (Fig. 2).

The basic computing element is a single unit, or "neuron," which computes a weighted sum of its inputs and then passes this sum through a nonlinear activation function—most commonly the rectified linear unit, or, "ReLU," which outputs all negative input as zero, and all positive input without transformation. These outputs are then treated as activations for the next layer of neurons. All mathematical operations are differentiable, and the weights between computing units are model parameters. This means that networks can be trained to make better predictions using stochastic gradient descent, in which the derivative of each weight can be calculated with respect to the output error, and its value updated accordingly over a number of iterations. For a more detailed review of the elements of supervised learning and deep neural networks, see Ref. 38.

These principles form the basis of the deep learning paradigm. However, they leave unspecified the exact connectivity and arrangement of layers of computing units. Discovering the connectivity, architecture, and objective function of a network that provides the right inductive biases to allow networks to perform well is a key part of deep learning research. The special operation that gives CNNs their name, convolution, is one such inductive bias. Typically, the lower layers of CNNs comprise many local feature detectors, each of which is repeated many times across the two-dimensional input. Mathematically, this corresponds to convolution: sliding a candidate visual pattern across a set of activations from the layer below, outputting a "feature map" that shows how strongly the pattern was detected recorded at each location (Fig. 2). The pattern might correspond to an oriented Gabor filter, a surface, or even a face.[17,39,40] In a typical CNN, many convolutional filters would be passed across activation values at each layer, outputting a stack of feature maps. The key to solving the classification problem is to learn the weights of these convolutional filters—that is, the patterns that the filters are detecting—from the training data.

There are a number of other important layer types in modern CNNs. For classification, the exact location of features (or combinations of features) is often less important than their presence—a property termed "local translation invariance." This allows a spatial down-sampling that is partially achieved by another special layer known as a "pooling layer," which contains neurons that simply output the average or maximum of a patch of input values. Another major concern when training learning algorithms with many weights is the problem of overfitting: the

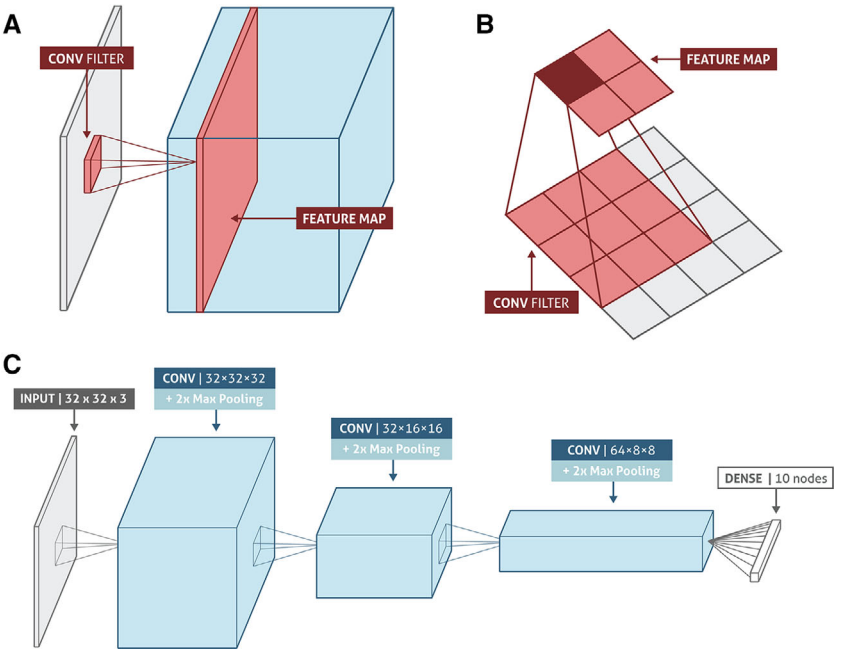

Figure 2. **A basic CNN architecture used for image classification. (A and B) Convolutional filters are moved across the activations of layers below, outputting feature maps. (C) A typical CNN architecture, based on AlexNet.[1] From a 32 × 32 RGB image, the first convolutional block learns weights for 32 feature maps, followed by two max pooling layers. The output is a vector of category probabilities; the category with the maximum of these values is taken as the output label. For computer vision tasks, the filters that are learned in the first few layers typically correspond to more general, low-level image features, such as oriented Gabors and colored blobs. The deeper layers tend to correspond to more task-specific, high-level features, such as faces or human or animal figures.[17,39,40]**

algorithm learns a solution that will not generalize well because it is too specific to the training set at hand. In machine learning, the statistical technique of "regularization" is used to protect against this phenomenon, by either including a term in the overall loss function that penalizes very large weights, and therefore extreme solutions, or by including regularization layers within the network. "Dropout" layers, which probabilistically drop the output of some units during training,[41,42] are often used for this purpose, meaning the network is forced not to rely on very specific coactivation patterns, but instead seek representations that are tolerant to noise and better distributed across nodes.

In the last few layers of the network, fully connected feedforward layers are often implemented. This gives extra capacity for representational flexibility and cross-class comparison. Finally, the classification layers consist of a linear decision rule for each category, converted into a probability through the softmax function:

$$p(Category\ i|\mathbf{x}) = \frac{\exp\{\boldsymbol{x} \cdot \boldsymbol{w}_i\}}{\sum_{i'} \exp\{\mathbf{x} \cdot \mathbf{w}_{i'}\}},$$

where $\mathbf{x}$ is the input representation from the final layer and $\mathbf{w}_i$ is the weight vector for category i. During training, these output probabilities are compared to the known output label, typically using the cross-entropy loss function, and the network backpropagates the error between these as the supervised learning signal. The description we have presented here characterizes the prototypical or perhaps archetypal CNN, and over the last decade, there have been many successful modifications of this architectural framework and training regime—for example, allowing lower layers to directly connect to higher ones.[43] For reviews of more recent developments, see Refs. 38, 44, and 45.

Viewed altogether, the effect of passing an image signal through a CNN is to learn the nonlinear

transformation of the input that supports the best linear separation between classes, with the aim of learning a mapping that will generalize well to other, related stimuli. The feedforward connections between stacks of layers mean that at each successive layer, increasingly abstract stimulus representations are formed. As we shall see below, all of these features make CNNs a promising candidate for extending cognitive models of similarity and categorizations to new stimulus domains.

Integrating CNN representations into models of higher cognition

One success from the study of higher-level cognition has been the development of a range of high-precision mathematical models for various aspects of human behavior,[46] and in particular a well-explored set of these models exists for the fundamental cognitive phenomena of similarity and categorization.[47–49] Although these modeling frameworks have driven an impressive range of theoretical and empirical developments, they have been most successful in a narrow range of cognitive settings—namely, human behavior in highly controlled experimental paradigms involving simple artificial stimuli.[50,51] This focus has been motivated by sensible aims. First, using simple and user-defined stimuli facilitates the general scientific method of isolating well-defined aspects of behavior and physiology, in order to evaluate theories and computational models at both the mechanistic and functional levels. Second, most models of higher-level cognition require as input stimulus representations that must "stand in" for inaccessible mental ones, and simple artificial stimuli can often be represented in a straightforward manner. Using CNN representations to extend these models to naturalistic stimuli more representative of the complex visual world is, then, an obvious and attractive target.

The classic technique for deriving representations used in cognitive models is multidimensional scaling (MDS), which seeks a spatial representation such that the distance between stimuli is inversely related to pairwise stimulus comparisons, such as similarity judgments.[52,53] The recovered dimensions are often found to be highly intuitive, and representations of stimuli, therefore, psychologically meaningful. However, MDS has two key limitations. The first is that it requires similarity judgments between *pairs* of stimuli, meaning data collection rapidly becomes unachievable as the desired stimulus set grows. The second is that there is no mechanism to generalize the inferred representation to novel items, which usually come *without* an accompanying set of similarity judgments with what we have already observed. Partially as a means of circumventing these limitations, studies using larger stimulus sets have designed simple artificial stimuli that can be directly represented using a low number of obvious perceptual dimensions of variation.[54] On one hand, these approaches result in straightforward stimulus representation and precise manipulation of participants' knowledge of stimuli—both of which allow experimental design and analysis to be focused on probing the difference between candidate modeling strategies. On the other hand, this means that the stimuli that have been used are typically far removed from the types that are used for reasoning about using our natural cognitive abilities and that most stimulate our desire to model the mind (Fig. 3).

The hierarchy of representations learned by CNNs for complex naturalistic inputs can help to close this gap. These representations are the outputs of intermediate functions in a composition (i.e., hidden layers), and thus constitute models of representation learning—although often only implicitly in service of the training objective such as classification, and not necessarily designed to mimic human perception. Further, because these networks are able to learn from millions of data points, and because they are validated on held out data, they generalize broadly, and can be used to characterize the content of diverse images. For example, CNNs trained exclusively to classify naturalistic, uncurated digital photographs of objects constitute impressive explanations of human shape sensitivity for novel, artificial shape stimuli that are devoid of context.[55] Taken together, these properties provide a compelling *a priori* argument that training such networks to solve human-relevant tasks over large datasets of naturalistic stimuli may produce general purpose representations, not unlike human ones, that can be used as a basis for cognitive modeling. However, just how similar these representations are to human psychological representations, and the nature of their discrepancies, is only recently coming into focus.

A first empirical indication that CNN representations might be of cognitive interest in this context came from a study by Lake *et al.*, who examined how

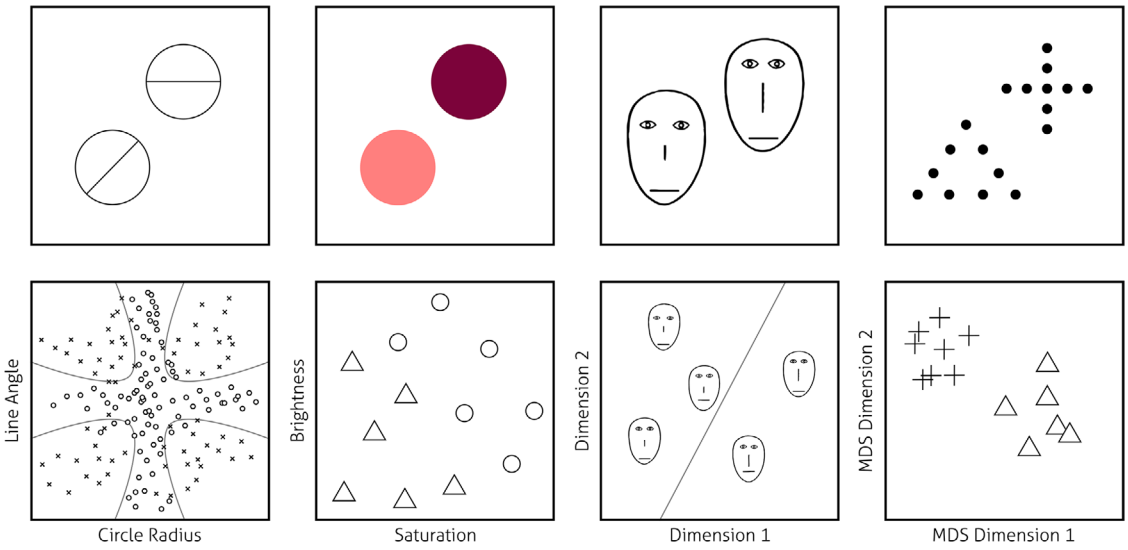

Figure 3. **Representative stimuli from seminal psychological studies of categorization. Top row: typical artificial stimuli representative of those used in traditional studies of cognition. Bottom row: the mathematical representation of these stimuli that are input into cognitive models. Reproduced from Ref. 50.**

well CNN activations at different layers predicted human category typicality ratings on novel images from eight ImageNet categories.[56] They tested three CNNs pretrained on ImageNet against the state-of-the-art classifier from computer vision prior to CNNs (scale-invariant feature transform, "SIFT" features[57] derived from ImageNet, followed by support vector machine, "SVM," classifiers).[58] The best predictive performance came from the final classification layer, where the average category activations from CNNs correlated very highly with human typicality ratings. The SIFT+SVM baseline, by contrast, scored poorly. Furthermore, for the categories where the CNNs performed poorly, there was some indication that this was due to the training image distribution from ImageNet being particularly skewed, and not as representative of the real world. Finally, the predictive ability of CNNs decreased monotonically with distance from this final layer, with the shallowest layers uncorrelated to the human judgments, indicating that, like human learners,[59] CNNs learn to use complex and category-specific features for natural stimuli as the basis of category typicality judgments.

In the remainder of this section, we summarize more recent results evaluating CNN representations in models of similarity and categorization. In general, we find that while CNN representations perform surprisingly well in capturing human behavior when used in conjunction with traditional cognitive models, they can always be made better by simple mathematical transformations. When related to each other, these methods constitute the beginnings of a modern toolkit for representing more naturalistic stimuli, based on data and theory from the cognitive sciences, statistics, and computer science.

Similarity

Judgments of stimulus similarity have long been of interest to psychologists, given both their intuitive importance to numerous cognitive processes[60] and their apparent correspondence to law-like human generalization behavior.[61] However, their most enduring influence in the field comes from their utility in revealing the structure of *psychological representations*—the comparatively high-level internal representations of external stimuli that are initially processed as raw sensory data. The precise structure of these representations is important to understand, because it shapes the downstream inferences that they support. For example, representing animals as a function of their relative size and color may lead to considerably different generalizations than representing them via their muscle mass and position on the food chain.

Similarity judgments provide strong clues as to which representations people might be employing, which are otherwise not directly observable. Moreover, compared to measurements of neural activity in visual brain areas, human similarity judgments exhibit additional structure and comparably little measurement noise.[62] Previous work often makes the assumption that similarity is a relatively simple metric that operates over pairs of stimuli in the representational structure (e.g., Euclidean distance between points in a metric space or Hamming distance for nodes in a hierarchy). Holding a given metric fixed, algorithms, such as multidimensional scaling, agglomerative clustering, and additive clustering, allow researchers to infer continuous and discrete feature representations that best explain patterns of observed human similarity judgments.[63] Applying these methods has been highly productive, often revealing crucial latent mental structure in sets of stimuli, and providing relatively unbiased candidates for stimulus representations that help explain downstream cognitive processes. Since these methods infer representations directly, they do not require a model of representation learning (i.e., an explanation of how the representations came to be, or why). The advantage of this approach is that it allows researchers to immediately begin to further study the processes that operate over these representations; but it is also a disadvantage, because it lacks exactly that level of explanation.

To begin to evaluate the precise correspondence between representations from CNNs and humans, Peterson *et al.* compared similarity scores (inner products) for pairs of images in the hidden-layer feature space of a CNN to average pairwise similarity ratings from people.[64] Across six sets of 120 images (720 in total) and over more than 400,000 total judgments, CNN representations were found to give a much better account of human behavior than traditional computer vision methods, capturing approximately half of the explainable variance in similarity ratings for five out of six image sets. While giving a reasonably good off-the-shelf fit, what seemed to be missing in the CNN representations was the taxonomic structure humans employ (e.g., making a pronounced distinction between primates and nonprimates in the case of images of animals), as revealed by hierarchical clustering (Fig. 4). Next, Peterson *et al.* showed that learning a set of dimensional weights that are applied to the inner product calculation in the CNN feature space produced similarities that captured nearly all of the remaining explainable variance in human judgments in most cases.[64] This formulation follows one long tradition in cognitive psychology of modeling similarity as the result of a comparison of objects in terms of a set of features,[65–67] expressed in matrix-vector notation as follows:

$$S(x, y) = \mathbf{x}^{\mathrm{T}} \mathbf{W} \mathbf{y},$$

where $\mathbf{x}$ and $\mathbf{y}$ are vectors of features associated with two stimuli, x and y, and $\mathbf{W}$ is a diagonal matrix that serves to weight each dimension in the comparison. Learning these weight parameters can be thought of as improving the similarity metric that operates over these representations in a way that better mimics human judgments—in particular, applying attentional weights to each feature. When these weights are constrained to be non-negative, it can also be thought of as a fixed linear transformation (axis-aligned scaling) of the original feature space, producing a modestly altered representation. The more human-like similarities obtained through this method also largely reproduced the latent hierarchical structure that was not previously observed in the CNN representation (Fig. 4). Since this method was validated on out-of-sample image pairs, it provides one of the first generalizable formulas for automatically producing human-like psychological representations for arbitrary naturalistic images: correct the CNN representation once, then apply the learned transformation to all future images that are input to the network.

Several subsequent studies have extended this method to a wider range of mathematical transformations, and in each case shown that doing so allows CNN representations to afford better and more interpretable models of human behavior. Attarian *et al.* showed the fit to human judgments can be improved further by allowing the transformation matrix, $\mathbf{W}$, to take increasingly unconstrained forms.[68] They applied these transformations to a dataset of similarity judgments over natural images of birds[69] that take a slightly different form from the absolute ratings described in the studies above: given a query image, participants must answer which of two (or more) reference images is most similar. Using the nonexponentiated Luce–Shepard choice rule,[52,70] the probability of

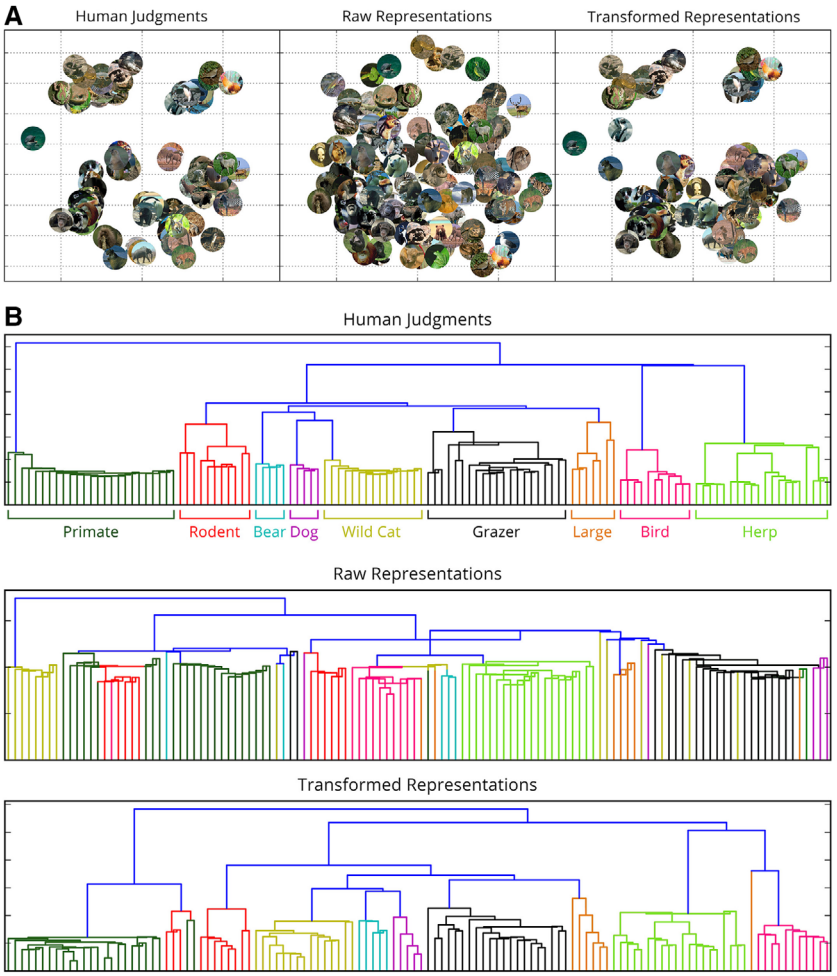

Figure 4. Transforming CNN representations using similarity judgments. (A) Representations of images derived from human similarity judgments using MDS exhibit meaningful variation and segregation (left panel). Using MDS to examine the similarity structure of raw CNN representations shows they fail to capture these relationships (center panel). Improving the fit of CNN representations to human similarity judgments recovers this structure (right panel). (B) Dendrograms of image representations also display meaningful hierarchical categorical structure (top panel) that is not present in raw CNN representations (middle panel) but that is recovered by modifying them using the learned similarity transformation (bottom panel). Reproduced from Ref. 64.

any single judgment becomes as follows:

$$p(y \text{ is more similar to } x|x,\ y,\ z) = \frac{S(x,y)}{S(x,y)+S(x,z)}.$$

Attarian *et al.* modeled each similarity comparison as follows:

$$S(x,y) = f(\mathbf{x})^{\mathrm{T}}\,\mathbf{W}\,f(\mathbf{y}),$$

where $\mathbf{x}$ and $\mathbf{y}$ are CNN representations of images x and y, and $f(\cdot)$ is a linear function that acts to reduce their dimensionality. The authors use principal components analysis (PCA) for this compression, such that the transformation matrix acts on the compressed space of k dimensions (i.e., $\mathbf{W} \in \mathrm{R}^{k\times k}$).

By choosing different constraints on $\mathbf{W}$, Attarian *et al.* allowed a range of linear transformations on the compressed CNN representations.[68] They compare the untransformed but compressed CNN representations ($\mathbf{W} = \mathrm{I}$), a dilation of compressed representations encoded in non-negative diagonal matrix characterized by a vector of diagonal components, $\mathbf{w}$, with and without regularization ($\mathbf{W} = diag(|\mathbf{w}|)$), a symmetric matrix ($\mathbf{W}_{ij} = \mathbf{W}_{ji}$), and a full, unconstrained matrix. In general, they

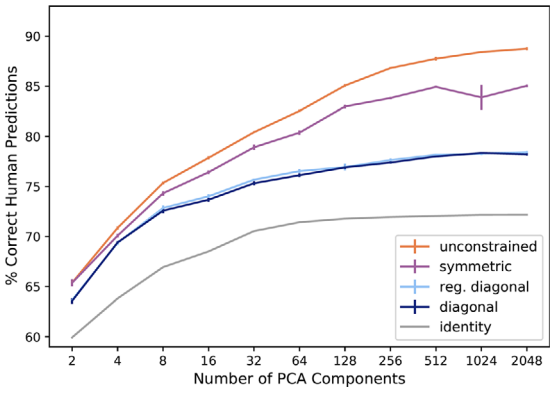

Figure 5. **Increasing the flexibility of the linear transformation of CNN representations improves fit to human similarity judgments. As the constraints on the transformation matrix are relaxed (see legend, bottom to top), the accuracy of model predictions increases. Increasing the number of principal components used to represent the compressed CNN representations also improves performance and widens the gaps between model subtypes. Error bars represent ± 1 SEM over five cross-validation folds. Reproduced from Ref. 68.**

find that each of these successive relaxations allows for a closer fit to the human judgments (Fig. 5), replicating and extending the findings in Ref. 64. Within these data are two other findings of note. The first is that regularization did not significantly improve the performance of the dilation, and the more complex models performed well on held-out triples. This indicates these transformations do not lead to overfitting despite large ratios of parameters to judgments—perhaps because of the functional restrictions to linear transformations. The second is that allowing similarity judgments to be asymmetric significantly improves the model. This is in line with the same finding from classical debates about the similarity in the psychological literature.[65]

Jha *et al.* proposed a model that jointly reduces the dimensionality of CNN representations *and* allows for more complicated linear transformations by using a bottleneck layer.[71] This lower-dimensional, fully connected linear layer is added on top of the final representational layer of a pre-trained CNN. It serves to learn a projection of images into a lower dimensional space such that the similarity between two images is given by the inner product of these lower dimensional and transformed representations. As every component is differentiable, it can be trained by backpropagation using similarity judgments themselves, allowing the dimensionality reduction to preserve information strictly relevant to similarity comparisons. This is in contrast to PCA, which aims to capture the variance of the original data in an unsupervised manner. In the context of the similarity transformation above, we can think of this as transforming and reducing the CNN representation space according to some $k \times d$ dimensional matrix, $\mathbf{V}$, where d is the dimensionality of the raw CNN representation (i.e., $\mathbf{V} \in \mathrm{R}^{k \times d}$).

Using the datasets originally given in Ref. 64, Jha *et al.* found that surprisingly few bottleneck dimensions are needed to offer a good approximation of human similarity judgments.[71] As the number of bottleneck dimensions increases, performance approaches that given by the full transformation over the 2^{12}-dimensional space of the original CNN representation. However, this increase is not linear, and an elbow to performance benefits is seen at around 2^4–2^6 dimensions (Fig. 6, top left). The PCA projections almost always perform worse than the bottleneck approach, corroborating the idea that the information relevant to similarity judgments is an important but not exclusive component of higher-level CNN representations (Fig. 6, top right).

Varying the size of the bottleneck layer offers the opportunity to explore what information from the CNN representations is used to model similarity judgments at different levels of granularity (Fig. 6, middle). Jha *et al.* found the principal components of the bottleneck representations reflected intuitive dimensions of variation in each original dataset, for example, separating images based on broad and continuous categorical distinctions, such as the animacy of objects, or the number of wheels.[71] Dendrograms constructed from bottleneck representations revealed that increasing the size of the bottleneck introduces finer levels of distinction (Fig. 6, bottom), mirroring the hierarchical aspect of clustering observed in human cognition and that found by parallel methods for the interpretation of CNN representations.[72]

One way to view the methods detailed above is as conducting fine tuning and transfer learning.[39] CNNs are first trained on large naturalistic image datasets, in order to ensure they learn relevant and general image features and do not overfit, and then these features are modified using similarity judgments to better highlight or extract the statistical information humans rely on when

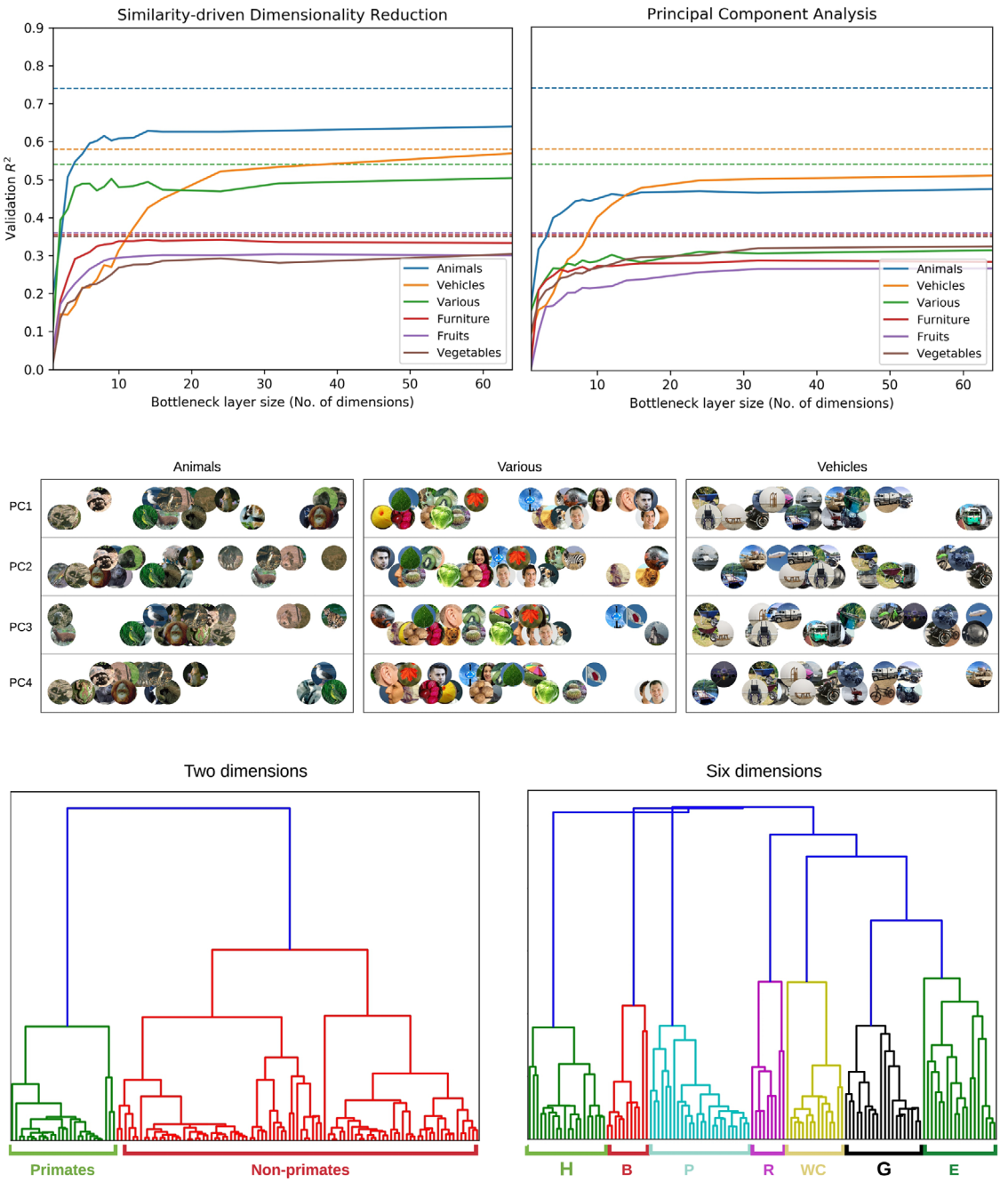

Figure 6. Exploring the effect of dimensionality reduction on modeling similarity judgments. Top row: CNN representations for a number of image datasets were reduced by using similarity judgments (left) or PCA (right). Performance using all representations and a simple dimensional reweighting are shown as a dashed line for each dataset. Middle row: fixing a small bottleneck size and projecting bottleneck representations using PCA allows interpretation of the information being encoded by the network and extracted for the similarity comparison. Bottom row: dendrograms for the animal dataset based on representations from bottleneck sizes of two and six show the CNN representation, and reduction captures similarity information according in a hierarchical manner. H, herps; B, birds; P, primates; R, rodents; WC, wild cats; G, grazers; E, dogs, bears, and large animals. Reproduced from Ref. 71.

making the same types of comparison. A second set of methods for refining CNNs in this manner has adapted pretrained CNNs to model MDS coordinates for images, derived from similarity judgments.[73–75] If the CNN is able to learn a faithful and psychologically meaningful MDS embedding, it can then be used to generate approximate coordinates for potentially infinite numbers of novel stimuli, overcoming the inherent lack of generalization from traditional MDS. In two studies, Sanders and Nosofsky[73,75] examined similarity judgments on a dataset comprising 360 images of subtypes of igneous, metamorphic, and sedimentary rocks.[76] First, they pretrained models on ImageNet to derive broadly generalizable base features and prevent overfitting to their own smaller dataset. Then, they generated a set of MDS coordinates for a subset of geological images, replaced the uppermost layers of CNNs with unlearned weights, and restarted training using the MDS co-ordinates as output targets for each image. If only the new layers are trained, this is known as "transfer learning"; if all layers are trained with a small learning rate, "fine-tuning."[39] After training, their CNNs produced approximate MDS coordinates, and in both cases, the authors find a reasonably high correlation between the CNN- and human-derived MDS coordinates for a held out set of images. In the second study, Sanders and Nosofsky assessed the generalizability of the raw MDS space and found that the majority of MDS dimensions had a high degree of subjective and objective correspondence with a second MDS space for a novel set of images from the same geological categories.[75] The remaining dimensions, however, were less easy to relate and exhibited lower correlation scores. The CNN-MDS approximations also showed reasonably high correlations with the "true" MDS coordinates over the four dimensions that were best preserved between original and novel image sets. The approximations were, however, quite poor for the remaining four.

The methods outlined above have been able to use behavioral data related to similarity to fine-tune CNN representations such that they better reflect human similarity structure. Peterson *et al.* investigated whether CNNs could be fine-tuned in this manner simply by using more general category labels.[77] The ImageNet labels that CNNs are normally trained on correspond to Rosch's "subordinate" level.[78,79] These are relatively specific, comprising, for example, breeds of dog (e.g., "dalmatian"). Peterson *et al.* trained CNNs on a subset of these labels for which the super-class label was also provided,[80] labels that corresponded to Rosch's "basic" level (e.g., "dog").[77] Forcing the CNN to learn a transformation that grouped stimuli according to both levels of this hierarchy was enough to lead to representations that both captured human similarity judgments better and exhibited much more structured and human-like taxonomic hierarchies. The authors also found that CNNs that were trained or supplemented with basic-level supervision provided a stronger match to the "basic-level bias" described in human generalization, so called because given a new label and an associated image, people will tend to generalize this label to other images at this level of description.[59,79] As with recovering the shape bias described in Ref. 55, ensuring that CNNs reproduce such biases is yet another avenue for assessing and improving their correspondence with humans.

The above studies provide evidence that, beyond comparable task performance, CNNs themselves may provide a useful source of image features that can be used to capture aspects of human psychological representations of naturalistic images. They also demonstrate that cognitively motivated strategies for modifying these representations can lead to better accounts of human behavior. Developing these strategies and integrating them more directly with the feature learning stage are important next steps for these frameworks, as well as examining important failure cases, in which the abstract correspondences between stimuli that humans identify are not well captured by CNNs.[81,82]

Categorization

A promising second candidate for the direct integration of CNN representations is the study of categorization—perhaps the most fundamental cognitive phenomenon following from similarity. The problem of categorization can be formulated as deciding how to assign a novel stimulus to a new or existing category, and it makes intuitive sense that this is in turn related its similarity with previously encountered category members. Although it seems like something has been lost—now only a subset of stimuli is being considered for any given comparison—categories add structure that

allows us to model within- and between-category interactions, and therefore a richer and more precise range of generalizations. Categories form the basis of most cognitive frameworks for understanding the brain and mind, and virtually all computational paradigms begin with a definition or explanation of them.[83]

Theories of categorization began with Plato and Aristotle, who posited that categories were best represented by a single ideal form or all previously encountered members, respectively.[84,85] More recently, the study of categorization has tracked the development of the cognitive sciences in general, beginning with the idea that people use rules or definitions to categorize stimuli.[86,87] However intuitive this idea seems, it was challenged in the 1970s by the work of Eleanor Rosch and colleagues, who showed that natural categories often lack central, defining features and exhibit graded category structure better characterized by a relation of family resemblance.[88–90] Strong evidence in support of this view came from a number of empirical findings, in which a stimulus could be more or less "typical,"[59,79] *unseen* prototypical stimuli were preferred to other unseen stimuli from the category during recognition tasks,[91] and that some levels of label (e.g., the "subordinate" level) worked better than others in explaining category judgments.[78]

Up until recently, the role of cognitive modeling has been to propose how the similarity between a stimulus and category should be computed in order to capture these behavioral effects, both in terms of the representation of existing category members and the formulation of the comparison itself. Two influential modeling strategies have been developed in this context. The first, known as a *prototype* model, assumes that categories are represented by their average or central tendency.[92] The second, known as an *exemplar* model, assumes that categories are represented by all of their known members.[93,94] These strategies can be unified by identifying the computational task of categorization as probability density estimation,[48] and in particular as inferring the probability a participant will choose a particular category label given an image. This more precise framing led to the fine-grained exploration of model predictions in laboratory settings, using sets of artificial stimuli that were designed to differentiate model performance, with exemplar models often winning out.[92–105] Since then, extrapolating the full probabilistic reframing of categorization has resulted in a number of other statistical strategies that can account for human behavior in a wider range of settings, such as learning mixture density estimators composed of a number of (sub)prototypes,[106,107] sharing exemplars across categories,[49] and allowing the categorization strategy to adapt flexibly to the number of stimuli observed.[24,49]

Battleday *et al.*[50] investigated whether these models could be extended to more naturalistic settings by using CNN representations to examine their predictions over a wide range of natural images. They first collected over 500,000 human categorizations on 1000 images from each of the 10 categories in the test subset of CIFAR-10,[108] a behavioral dataset they call CIFAR-10H. For each image, they generated a range of machine learning representations (Fig. 7), including raw pixel representations, features derived from computer vision and engineering (histograms of oriented gradients; HOG), the activations of the latent layer of a generative adversarial network that uses convolutions (BiGAN), and the upper-layer activations of two CNNs—AlexNet[1] and DenseNet.[43] Finally, they generated predictions from a range of prototype and exemplar models using the machine learning representations. These strategies draw on a second influential tradition in modeling similarity: this time as an exponentially decreasing function of distance in some psychological space.[61] Prototype and exemplar models specify this similarity calculation mathematically, and convert it into a probability by relating it to the inverse distance to the category prototype or summed inverse distances to all category members, respectively:

$$S(y,\ C)_{Prototype} = \exp -d(\mathbf{y},\ \mu_c),$$

$$S(y,\ C)_{Exemplar} = \sum_{\mathbf{x}\in C} \exp -\beta d\,(\mathbf{y},\ \mathbf{x}),$$

where C is a category, y is a stimulus, $\mathbf{y}$ is the multidimensional representation of that stimulus, and $S(\cdot, \cdot)$ is our similarity function. This representation is compared with the category mean for prototype models or all known category members for exemplar models. The similarity function, S, is *additive*: if a category is represented by a vector of stimuli (as in the exemplar model), S computes the sum of

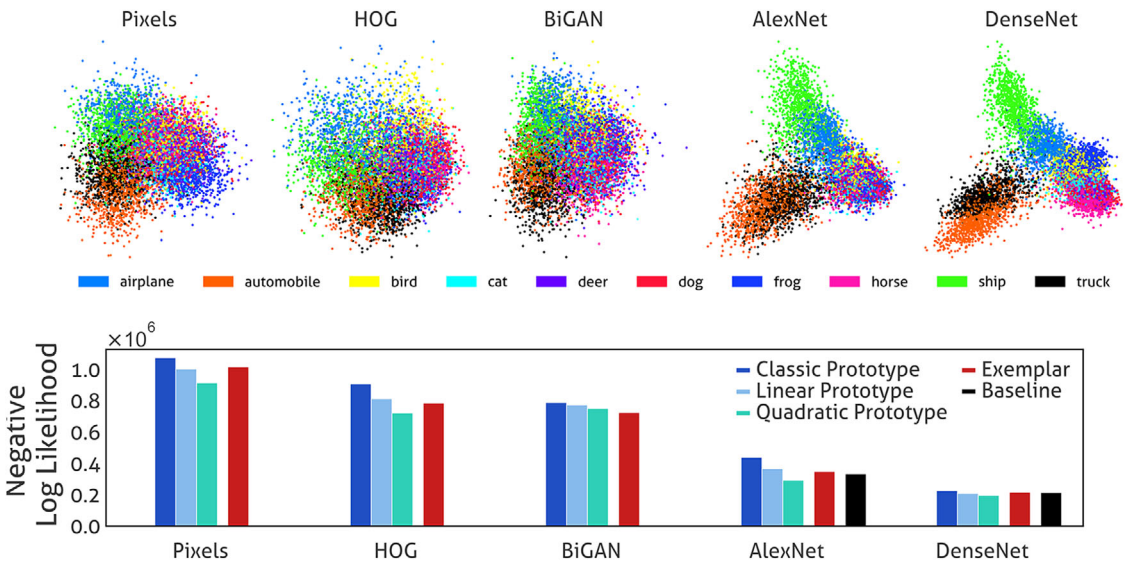

Figure 7. **The feature basis used for modeling categorizations of natural stimuli affects overall model performance more than categorization strategy. Top row: two-dimensional linear discriminant analysis projections of the representations from each computer vision method. The feature bases across the x-axis roughly track the development of computer vision: raw pixels, hand-engineered features (HOG), the latent space of a generative network that uses convolutions (BiGAN), and a basic (AlexNet) and more advanced (DenseNet) CNN. Bottom row: categorization models using different prototype and exemplar strategies were trained on each of these feature bases, with model flexibility being more obviously related to overall model performance than categorization strategy (i.e., prototype or exemplar). Baselines were provided by taking the softmax probabilities from the final CNN layer as the similarity measurement. Reproduced from Ref. 50.**

similarities between **y** and the representation, **x**, for each stimulus in *C*. Finally, the exemplar model contains a "specificity" parameter, β, that scales all distance calculations by the same amount prior to exponentiation. This acts to sharpen or lessen the influence of exemplars on subsequent category judgments, and therefore allow the model to controls for overall stimulus discriminability in the relevant psychological space.[94]

Battleday *et al.*[50] found that CNNs provided the best representational basis for modeling the human categorization judgments, outperforming deep unsupervised and traditional computer vision methods. Indeed, the choice of stimulus representation affected overall performance to a much greater extent than the choice of categorization model. This is particularly interesting given the main focus in categorization modeling has been the categorization strategy, whereas the unambiguous nature of simple artificial stimuli has allowed their representation to remain fixed. The loss of such neat factorization when analyzing naturalistic stimuli establishes a need—at least initially—to focus on the less well-integrated question of perceptual representation. When restricting analysis to the CNN representations, categorization models with more free parameters—and therefore more flexibility to transform CNN representations—performed the best, and better than competitive baselines. This indicates that the CNN representations also naturally contain latent information relevant to the kinds of features humans use to make category judgments, which can be further refined using simple linear and quadratic transformations. Another unexpected finding was that prototype and exemplar models performed roughly the same across a varied range of image representations, contrasting with what might be expected based on previous laboratory work in which exemplar models are needed to capture artificial categories with more complex structure.

There *are* known cases in which prototype and exemplar models make similar predictions, for example, if category representations are well-captured by simple Gaussians, but how these situations relate to high-dimensional representations of complex stimuli remains unclear. Battleday *et al.*[50] conducted a simulation study that investigated how the number of dimensions and training samples

affected the different modeling strategies. They found that while indeed in low-dimensional spaces exemplar models outperform prototype models, no such difference exists in high-dimensional representational spaces after training with very large numbers of stimuli. This is the regime in which CNN-based representations of naturalistic images allow us to operate. Further theoretical investigation is needed to frame the import of these dependencies, as well as to make precise the relationship between the size and nature of representational space.

A number of the studies investigating CNNs in the context of similarity outlined above have used categorization as a downstream application to test the utility of the representational spaces their methods derive. For example, Peterson *et al.* noted that people also found it much easier to learn novel categories of natural images when these were constructed based on the similarity between CNN representations rather than similarity between HOG/SIFT computer vision engineering features.[64] When these CNN representations had been transformed to better align with human similarity data, categories were even easier to learn. Second, Sanders and Nosofsky examined categorizations for a held-out set of data taken over 120 novel images of geological samples, taken from the same 30 categories that they used to train CNNs to approximate MDS coordinates (see above).[73] When a low-parameter exemplar model was supplied with the CNN-MDS representations, it was able to model these human categorizations very well. In a further study, Sanders and Nosofsky also tested an exemplar model on this categorization data, but this time compared its performance using a number of CNN representations, including a set modified to better reflect human (dis)similarity judgements.[75] They found the CNN-derived MDS-approximations allowed much better performance than the simple, off-the-shelf CNN image representations, and could be further improved by training the CNN to also predict "missing" dimensions, appended to the normal MDS co-ordinates.[76]

A convergent set of findings comes from the closely related field of object recognition. Annis *et al.*[109] showed that CNNs pretrained on ImageNet generated intuitive, correctly clustered, and robust representations of entirely novel experimental objects, including *Greebles*,[110] *Ziggerins*,[111] and *Sheinbugs*.[112] These more complex artificial stimuli exhibit naturalistic qualities and are difficult to represent by a few obvious dimensions or MDS, and yet ensure participants have no prior experience with the object category. Annis *et al.* then used these CNN representations to model a set of data[113] in which participants had to decide whether a second object, possibly rotated or scaled, was the same as a first, separated by a delay.[109] The similarity between the first and second objects was used as the input to a hierarchical Bayesian evidence accumulation model in order to model participants judgments and reaction times. As with the categorization studies above, Annis *et al.* found that using CNN representations together with the cognitive model outperformed a number of baselines, that the choice of CNN representation was the major determinant of performance, and that the best-performing model could be improved by modifying the CNN representations to also be robust to rotations. They also explored a range of methods for finding transformation, complementing the analysis conducted by Refs. 64 and 68 for similarity. See also Ref. 114, for another successful application of CNNs and evidence accumulation models, this time in modeling the decisions and reaction times of novices and experts in medical image classification.

There are a number of other studies that investigate aspects of the relationship between image categorization, cognitive models, and CNNs. As with category typicality judgments,[56] the predictive benefit of CNN representations for modeling categorizations was found to increase with increasing depth.[115] On the other hand, representations from shallower CNN layers provided the best basis for capturing visual biases observed in in particular pigeons classifying cardiograms,[116] suggesting a role for CNNs in comparative studies of visual categorization across other types of visual system.

Along with typicality and similarity, CNNs appear our best available source of representations for modeling behavior over naturalistic stimuli in another fundamental cognitive domain—categorization. In general, studies have found two approaches successful in improving these representations further: adapting them to better fit similarity judgments *before* applying categorization models, or by designing flexible categorization models that transform CNN representations further *during* parameter fitting. More generally, it appears the

role of stimulus feature learning can no longer be withheld from formal accounts of categorization over more complex types of stimuli and considering both together will be necessary for characterizing human category judgments. This echoes a long-accepted perspective in the machine learning literature and has been called for previously in the cognitive sciences.[117]

Improving deep neural networks using concepts from cognitive modeling

In the studies above, CNNs have been primarily used to provide representations of stimuli to which traditional models of higher cognition can then be applied. We have also seen that it has always been possible to make these representations better by transforming them to fit held-out behavioral data. In applying and modifying CNN representations for this new set of problems, a number of insights and models have been developed that are also valid to the machine learning and computer vision contexts in which CNNs were originally developed. In this section, we discuss how this knowledge can be integrated back into normal CNN training paradigms in order to make them better cognitive models, with the simple aim of improving their performance.

A central goal of the transformations above was to better emulate the similarity structure and graded category membership assumed by cognitive models and exhibited in human categorization data. This type of structure has been a central and organizing feature of categorization theory in general and has been found to widely apply to natural categories in general, including the types of stimuli CNNs are typically applied to.[59,79,88,89] In the field of natural image classification, however, emulating such structure is not normally viewed as a priority. Rather, the "top-one" accuracy of a classifier is prioritized, achieved through associating a "ground-truth" or "hard" label with each image that assigns it to a single category with no room for uncertainty; the CNN must learn a feature space that supports such *n*-ary discrimination. This modeling choice has until now been understandable—until recently, the challenge for classifiers was simply getting the most likely image label right. It is only now, when CNN accuracy is beginning to asymptote at near-human performance, that secondary features of CNN performance are beginning to come under scrutiny—for example, their poor generalization to out-of-training-distribution images, and their fragility in the face of adversarial attacks.[118–122] There has been broad appreciation that to tackle these challenges image classification models must be better able to deal with noise, and one means of achieving this is by endowing them with better perceptual models of the world. Indeed, there is some evidence that when human participants are trained to label stimuli with single categories based on stimulus features, as opposed to making predictive inferences about features or building a generative model of the data, they are less likely to learn the kinds of category abstraction that are thought to underlie graded typicality judgments.[56,123,124]

A prime motive for trying to integrate cognitively motivated category structure back into CNN learning paradigms is, then, that it may help provide a solution to this new frontier of problems. Peterson *et al.*[125] attempted to do so *indirectly* by training a range of CNNs to predict the human categorization judgments ("human" or "soft" labels) from the CIFAR-10H dataset. They found that these CNNs performed significantly better on a number of out-of-sample natural image sets than CNNs trained on CIFAR-10, while scoring the same on the hard-label validation set (Fig. 8, left and center). As test images were drawn from increasingly out-of-sample datasets (i.e., decreasingly similar), the benefit from using human labels increased, indicating these networks were learning more perceptually relevant category distributions. This was also reflected in their higher second-best accuracy scores: whether their second-best guess for an image corresponded to humans (Fig. 8, right). The authors also subjected these networks to various forms of adversarial attack and found that CNNs trained with human labels were much more resilient. It appeared as though training with human labels endowed networks with more tolerance to noise and more graceful degradation: exactly the current aims of the computer vision community, and those properties originally sought by early artificial intelligence researchers. Finally, Peterson *et al.*[125] showed that CNNs trained with human labels invariably performed better than alternative strategies, which either incorporate random label noise or train on convex combinations of image-label pairs.[126] This shows that the structure contained in human labels is helpful for classification beyond the regularization effects of adding training noise.

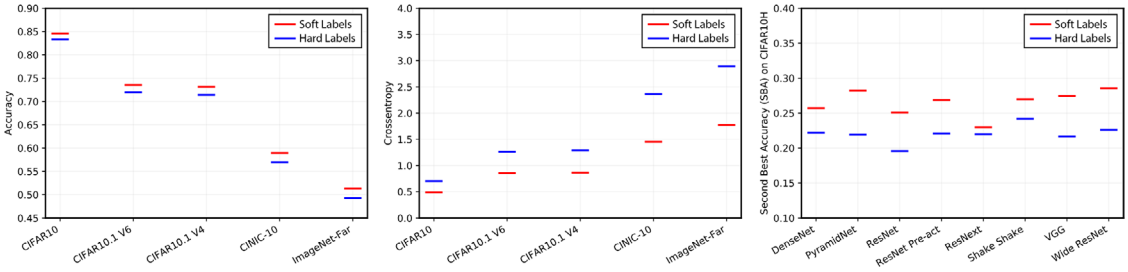

Figure 8. **Improving the generalization abilities of CNNs using human uncertainty. As test images come from increasingly out-of-training-sample distributions, CNNs trained on soft labels derived from human uncertainty increasingly outperform their traditional hard-label counterparts (in terms of accuracy and loss). The difference in distributional benefits to using hard labels is reflected in the consistent benefits regarding second-best accuracy scores. Reproduced from Ref. 125.**

A second area where cognitive models have inspired successful computer vision models is in the context of few-shot, zero-shot, and semisupervised learning—all variants of image classification tasks in which CNNs are trained with a much smaller subset of the data than normal. These reductions in the size and coverage of the training set means stronger or more suitable inductive biases must be built into CNNs in order to prevent overfitting and make successful generalizations from fewer data. Here, the strategies that have been successful *directly* incorporate graded category structure into their training paradigm by using probabilistic constraints on a CNN's classification function, many of which come directly from the range of categorization models reviewed above.

In each episode of a few-shot learning task, a classifier must identify the label of a query image given a support set of image-label pairs. Within the support set, there are a fixed number of examples from a subset of possible categories, and the classification algorithm must construct categorical knowledge from this small number of examples in order to identify which category the query image belongs to. The number of "shots" refers to the number of training examples per category, and the "way" refers to the number of categories: for example, "one-shot five-way" means identifying which of five possible category labels best applies to the query image based on one labeled example from each category.[127,128] By training over many of these episodes, a successful classification algorithm learns how to use the examples in the support sets to generalize categorical knowledge to the query example over many different combinations of images and categories.

In the prototypical networks (PNs) of Snell *et al.*, isotropic Gaussian category distributions with identity covariance matrices are used over final layer representations of CNNs in order to classify novel stimuli.[129] By representing categories in this manner, PNs reduce classification to the calculation of (Euclidean) distances between category prototypes, drawing on an equivalence between prototype models and Gaussian classifiers first described in Ref. 48. PNs use a standard CNN to transform images into deep representations in the regular manner, such that the CNN weights are shared across support set members and episodes. The category prototype for each episode is the empirical average of labeled category in the support set, and the role of the CNN during training is to learn the best transformation to support prototype construction. Simply adapting the final-layer comparison in this way led to state-of-the-art performance on few-shot classification over the Omniglot character recognition task,[127] mini-ImageNet,[128] and even for zero-shot learning—in which only a vector of metadata is given about categories, with no examples—on the Caltech UCSD bird dataset.[130]

Following the development of cognitive categorization models as increasingly complex strategies for probability density estimation, Scott *et al.* were able to improve on PNs by using a wider range of distributions to model category structure, an approach they call stochastic prototype embeddings (SPE).[131] Each image was still used to generate a deep CNN representation that would be combined to model a prototype over each of the categories in the support set. However, Scott *et al.* also learned an additional output representation for each image to model category variance, allowing for more

flexible categories that followed axis-aligned ellipsoidal Gaussian distributions. They also altered the prototype embedding function to be stochastic, using a probabilistic model that incorporates a global noise distribution over latent prototypes. These two modifications allowed the CNN to learn embeddings that supported more accurate and robust distributions more likely to handle uncertainty arising from, for example, out-of-distribution inputs due to measurement noise and label uncertainty due to overlapping classes. For one-shot classification tasks in low-dimensional embedding spaces on Omniglot, the SPE model outperformed the deterministic PN and achieved state-of-the-art results. On the other hand, when five training examples were used in each episode instead, or in higher dimensions, the uncertainty over prototypes proved less important, and the models performed similarly. Their most impressive finding was in the context of data corruption. On an N -digit version of MNIST,[132] the SPE consistently performed better than a number of baselines, including PNs, when input images were randomly corrupted, validating the idea that SPEs learn categorical structures that are more robust to perceptual noise.

Allen *et al.* allowed for even more complex category structure with infinite mixture prototypes (IMP), which model each category as a (possibly infinite) mixture of Gaussian distributions.[133] The authors built this model based on two key insights from cognitive models of categorization: first, the idea that complex categories might be best represented by more than one cluster—an interpolation toward exemplar models had been previously explored in cognitive science;[106,107] and second, the idea that the number of category clusters should be free to grow as needed and the nonparametric Bayesian machinery to support it.[24,49] Combining these insights led to a differentiable model for generating an arbitrary number of subprototypes for a given category, which could then be used for subsequent learning and classification tasks. For few-shot *character* classification tasks in Omniglot, the IMP performed similarly to PNs and nearest neighbors.

The equivalence between these methods is unsurprising, as distributions over characters are likely to take some unimodal form well captured by a single prototype. However, as the classes being modeled became more complex, the IMP performed much better. On five-shot five-way tasks from *mini*-ImageNet, which is based on natural image categories that are likely to have more complex and multimodal structure, the IMP achieved state-of-the-art results. The IMP also performed extremely well on ten-way ten-shot *alphabet* recognition on Omniglot, around 25% better than PNs. This is again because alphabets of individual characters likely constitute complex distributions, with multiple modes corresponding to clusters of character types and styles. Allen *et al.* demonstrated that learning these more complex distributions was also beneficial for transfer learning.[133] They showed that when the IMP is trained on alphabet recognition tasks and tested on character recognition tasks from *Omniglot*, it outperforms a PN trained just on character recognition tasks. This is a significant result, because it implies something extra about individual characters has been learned by modeling their distribution within a specific alphabet, and that the *inference* model can be responsible for capturing this—in this case by maintaining flexibility over clusters that can vary with the task instance at hand. Even more impressively, when particular characters are held out from training at the alphabet level, the IMP successfully generalized its knowledge to recognize them as part of the alphabet at test time, to a much greater extent than PNs. These positive transfer learning results were also found to hold for a tiered version of ImageNet,[134] where fine-grained and multimodal structure captured by the IMP from image distributions at the level of complex classes (i.e., when using supercategory labels) transferred well to image distributions under from lower-level category labels. In both cases, the PN trained on superclasses and tested on subclasses performs much worse. Because it uses a soft-clustering scheme that allows unlabeled examples to act as supports, Allen *et al.* were able to apply the IMP to semisupervised versions of Omniglot and *mini*-ImageNet, in which only a subset of support images come with training labels, and in learning the unsupervised regime inferred clusters of Omniglot. In the semisupervised tasks, the IMP achieved state-of-the-art performance, and in the unsupervised task, it provided meaningful clusters of characters.

The above studies are an encouraging sign that integrating extra information from learners with pressures to learn generative and predictive visual categories into CNNs through training objectives

and datasets is a promising route to strengthening their generalizability and robustness. When data are plentiful, in the context of normal image classification, adding this information indirectly—for example, in the form of augmented image labels—is likely to be the most valuable strategy, in order to best exploit the extremely flexible learning capacity of CNNs. When data are scarce, however, stronger inductive biases that can be directly incorporated into the classification function used by CNNs have led to better performance. Indeed, learning and making robust inferences from limited information is a hallmark of human cognition,[135] and so these are natural settings to consider integrating human-like constraints.

End-to-end cognitive models

So far, we have seen CNN representations extend traditional models of higher-level cognition to more ecologically valid stimuli, and cognitive considerations and constraints modify CNN training paradigms such that they become better performers. The remaining challenge is to bring these frameworks even closer together, and fully leverage the benefits of both approaches to build the next generation of cognitive models.

A main clue about how to do this comes from the investigations already outlined above. In these studies, the primary advantage of CNNs was their ability to learn rich representations of a diverse range of visual stimuli. This ability arises from their flexible, gradient-based training regimes, as well as the special functions of each layer, and these properties were not modified at all during their use in cognitive models or reapplication to computer vision tasks. Instead, there was a search over CNN representations to find the best support for a behavioral dataset, followed by investigating which of a range of cognitive models were able to best fit the data. From a cognitive perspective, this amounts to taking a two-step search, identifying the best combination for the representation of stimuli and the representation of higher-level cognitive structures such as categories. But, there is no reason to separate this process. Provided our cognitive models can be made differentiable, the same training regime can be applied to both CNN and cognitive parameters instead, allowing the CNN and cognitive components to jointly learn which features of images are necessary to support the best application of a cognitive model to human behavioral data.

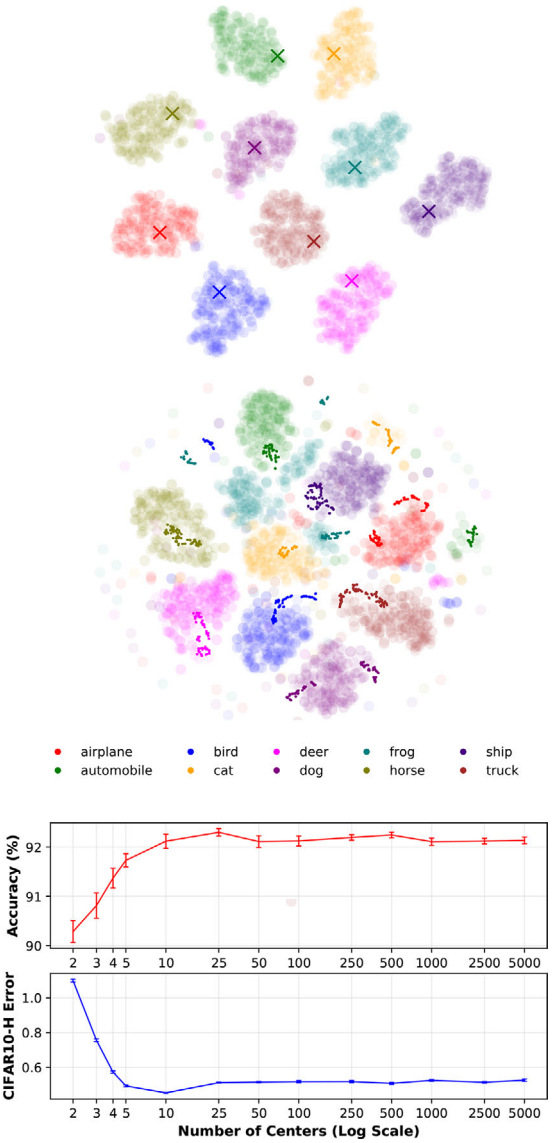

Figure 9. Deep categorization models learn category-specific stimulus embeddings. Top and middle rows: t-distributed stochastic neighbor embeddings[138] of representations from a deep prototype (top) and deep GMM with 25 centers (bottom), with locations of prototypes and subprototypes marked, respectively. Bottom row: performance in the GMM increases with the number of centers until around 10−25, then asymptotes. Reproduced from Ref. 136.

A first attempt in this direction was made by Singh *et al.*, for a range of the categorization models reviewed above.[136] Fortunately, the functional

forms of prototype, Gaussian-mixture (GMM), and exemplar categorization models can be made differentiable with only slight modifications, and are similar enough to normal CNN classifiers that they can simply replace them as the top layer of an otherwise normal CNN. In this setup, the location of the prototypes, subprototypes, and exemplars, respectively, are free parameters, which the models must learn from some training subset of labeled images. As usual, the role of the lower CNN layers is to learn the best feature representation to support these kinds of category comparison. Singh *et al.* trained a prototype model, GMMs with 2−2500 centers, and an exemplar model across two CNN architectures.[136] Their input data were the training subset of CIFAR-10: 50,000 image-label pairs, subdivided into 10 categories (i.e., image-hard-label pairs). They tested model generalization to the CIFAR-10 validation set: 10,000 image-hard-label pairs, as well as the CIFAR-10H dataset: 10,000 image-soft-label pairs (i.e., distributions derived from human categorizations). The first finding was that, again, choice of CNN architecture was the major determinant of performance and generalization. The second was that despite performing similarly on the hard label validation data, the cognitive models performed *much* better in predicting human categorizations than their CNN baselines. Varying the number of centers in the GMM revealed that the optimal number of centers—subprototypes—was around 25, which overall provided the best performance on human data (Fig. 9). Beyond modeling the human data well, the fact that these end-to-end models were able to outperform CNN baselines on human data even when trained exclusively with hard labels is an important validation of the functional forms arrived at by the cognitive psychology literature, as well as the development of categorization models beyond prototypes and exemplars to mixture models and beyond. It remains to be seen what further improvements arise when such models are *trained* on human behavioral data, in addition to being tested on it.

Conclusion

A major success from traditional studies of higher-level cognition was the development of a range of computational models that make high-precision predictions about fundamental human behaviors. To apply these models, it was necessary to develop statistical techniques to represent stimuli in a psychologically meaningful manner, which in turn limited their application to simple artificial stimuli. In this review, we have shown that we can use CNNs as the basis for a new set of statistical techniques that allow the representation of more naturalistic and varied stimuli. Indeed, these methods differ from techniques like MDS in a *quantitative*, rather than qualitative way. Both sets use a differentiable objective function to iteratively optimize the statistical fit of a multidimensional vector to a set of human decisions, as well as some degree of hyperparameter search (e.g., the number of dimensions in MDS *versus* layer size in a CNN). The main difference is, then, the sheer number of parameters involved in learning a representation that can be generalized to an arbitrary number of naturalistic stimuli. It makes sense that this process be constrained both by information about the distribution of these stimuli in the natural world and the structure of human knowledge regarding them, and we have highlighted how the CNN and cognitive components of more integrated models can be used to bring these sets of constraints closer together. Along with the fundamental scientific interest in extending analysis of behavior to a more complex and ecologically valid domain, as well as the general aim of increasing the precision of our efforts to model higher-level cognition, this new set of models has the added benefit of being directly applicable to the wealth of psychologically relevant data now available online.[137]

Acknowledgments

This work was conducted under Grant Number 1718550 from the National Science Foundation. R.M.B. drafted the manuscript. J.C.P. created the figures. T.L.G. conceived the paper design. All authors edited the manuscript. The authors would like to thank Thomas Palmeri and Jason Chow for their helpful comments in reviewing the manuscript.

Competing interests

The authors declare no competing interests.

References

1. Krizhevsky, A., I. Sutskever & G.E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* 1097–1105.

2. Baxter, J. 2000. A model of inductive bias learning. *J. Artif. Intell. Res.* **12:** 149–198.
3. Russakovsky, O. *et al.* 2015. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115:** 211–252.
4. Duta, I.C., L. Liu, F. Zhu & L. Shao. 2020. Pyramidal convolution: rethinking convolutional neural networks for visual recognition. *arXiv preprint arXiv:2006.11538*.
5. Lin, T.-Y. *et al.* 2014. Microsoft COCO: common objects in context. In European Conference on Computer Vision 740–755.
6. Qiao, S., L.-C. Chen & A. Yuille. 2020. DetectoRS: detecting objects with recursive feature pyramid and switchable atrous convolution. *arXiv preprint arXiv:2006.02334*.
7. Goyal, Y., T. Khot, D. Summers-Stay, *et al.* 2017. Making the V in VQA matter: elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 6904–6913.
8. Jiang, H., I. Misra, M. Rohrbach, *et al.* 2020. Defense of grid features for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 10267–10276.
9. Johnson, J. *et al.* 2017. CLEVR: a diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2901–2910.
10. Perez, E., F. Strub, H. De Vries, *et al.* 2017. Film: visual reasoning with a general conditioning layer. *arXiv preprint arXiv:1709.07871*.
11. Mnih, V. *et al.* 2015. Human-level control through deep reinforcement learning. *Nature* **518:** 529–533.
12. Kriegeskorte, N. 2015. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annu. Rev. Vis. Sci.* **1:** 417–446.
13. Khaligh-Razavi, S.-M., L. Henriksson, K. Kay & N. Kriegeskorte. 2017. Fixed versus mixed RSA: explaining visual representations by fixed and mixed feature sets from shallow and deep computational models. *J. Math. Psychol.* **76:** 184–197.
14. Bashivan, P., K. Kar & J.J. DiCarlo. 2019. Neural population control via deep image synthesis. *Science* **364:** eaav9436.
15. Hong, H., D.L. Yamins, N.J. Majaj & J.J. DiCarlo. 2016. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nat. Neurosci.* **19:** 613–622.
16. Agrawal, P., D. Stansbury, J. Malik & J.L. Gallant. 2014. Pixels to voxels: modeling visual representation in the human brain. *arXiv preprint arXiv:1407.5104*.
17. Güçlü, U. & M.A. van Gerven. 2015. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* **35:** 10005–10014.
18. Khaligh-Razavi, S.-M. & N. Kriegeskorte. 2014. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* **10:** e1003915.
19. Schrimpf, M. *et al.* 2018. Brain-Score: which artificial neural network for object recognition is most brain-like? https://doi.org/10.1101/407007
20. Yamins, D.L. *et al.* 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. USA* **111:** 8619–8624.
21. Cadieu, C.F. *et al.* 2013. The neural representation benchmark and its evaluation on brain and machine. *arXiv preprint arXiv:1301.3530*.
22. Cadieu, C.F. *et al.* 2014. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Comput. Biol.* **10:** e1003963.
23. Marr, D. 1982. *Vision*. San Francisco, CA: W. H. Freeman.
24. Anderson, J.R. 1991. The adaptive nature of human categorization. *Psychol. Rev.* **98:** 409–429.
25. McCulloch, W.S. & W. Pitts. 1943. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **5:** 115–133.
26. Rosenblatt, F. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev*. **65:** 386–408.
27. Hebb, D.O. 1949. *The Organization of Behavior: A Neuropsychological Theory*. Wiley.
28. Hodgkin, A.L. & A.F. Huxley. 1952. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.* **117:** 500–544.
29. Rumelhart, D.E., G.E. Hinton & R.J. Wilson. 1986. Learning representations by back-propagating errors. *Nature* **323:** 533–536.
30. Elman, J.L. 1990. Finding structure in time. *Cogn. Sci.* **14:** 179–211.
31. Rumelhart, D. & J. McClelland. 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge, MA: MIT Press.
32. LeCun, Y. *et al.* 1989. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1:** 541–551.
33. Riesenhuber, M. & T. Poggio. 1999. Hierarchical models of object recognition in cortex. *Nat. Neurosci.* **2:** 1019–1025.
34. Fukushima, K. & S. Miyake. 1982. *Competition and Cooperation in Neural Nets*. Springer.
35. Denker, J.S. *et al.* 1989. Neural network recognizer for hand-written zip code digits. In *Advances in Neural Information Processing Systems* 323–331.
36. Mozer, M.C. 1987. *Early Parallel Processing in Reading: A Connectionist Approach*. Lawrence Erlbaum Associates, Inc.
37. Radford, A. *et al.* 2019. Language models are unsupervised multitask learners. *OpenAI blog* **1:** 9.
38. LeCun, Y., Y. Bengio & G. Hinton. 2015. Deep learning. *Nature* **521:** 436–444.
39. Yosinski, J., J. Clune, Y. Bengio & H. Lipson. 2014. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems* 3320–3328.
40. Zeiler, M.D. & R. Fergus. 2014. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision* 818–833.
41. Srivastava, N., G. Hinton, A. Krizhevsky, *et al.* 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15:** 1929–1958.

42. Hinton, G.E., N. Srivastava, A. Krizhevsky, *et al.* 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.
43. Huang, G., Z. Liu & K.Q. Weinberger. 2016. Densely connected convolutional networks. *arXiv preprint arXiv: 1608.06993*.
44. Rawat, W. & Z. Wang. 2017. Deep convolutional neural networks for image classification: a comprehensive review. *Neural Comput.* **29:** 2352–2449.
45. Khan, A., A. Sohail, U. Zahoora & A.S. Qureshi. 2020. A survey of the recent architectures of deep convolutional neural networks. *Artif. Intell. Rev.* **53:** 5455–5516.
46. Sun, R. 2008. *The Cambridge Handbook of Computational Psychology*. Cambridge University Press.
47. Griffiths, T.L. & J.B. Tenenbaum. 2001. Randomness and coincidences: reconciling intuition and probability theory. In *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society*. J.D. Moore & K. Stenning, Eds.: 370–375. Mahwah, NJ & London, UK: Lawrence Erlbaum Associates.
48. Ashby, F.G. & L.A. Alfonso-Reese. 1995. Categorization as probability density estimation. *J. Math. Psychol.* **39:** 216–233.
49. Canini, K.R. & T.L. Griffiths. 2011. A nonparametric Bayesian model of multi-level category learning. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*.
50. Battleday, R.M., J.C. Peterson & T.L. Griffiths. 2020. Capturing human categorization of natural images by combining deep networks and cognitive models. *Nat. Commun.* **11:** 5418.
51. Nosofsky, R.M. 1992. Similarity scaling and cognitive process models. *Annu. Rev. Psychol.* **43:** 25–53.
52. Shepard, R.N. 1957. Stimulus and response generalization: a stochastic model relating generalization to distance in psychological space. *Psychometrika* **22:** 325–345.
53. Kruskal, J.B. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **29:** 1–27.
54. Ashby, F.G., G. Boynton & W.W. Lee. 1994. Categorization response time with multidimensional stimuli. *Percept. Psychophys.* **55:** 11–27.
55. Kubilius, J., S. Bracci & H.P. Op de Beeck. 2016. Deep neural networks as a computational model for human shape sensitivity. *PLoS Comput. Biol.* **12:** 1–26.
56. Lake, B.M., W. Zaremba, R. Fergus & T.M. Gureckis. 2015. Deep neural networks predict category typicality ratings for images. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*.
57. Lowe, D.G. 1999. Object recognition from local scale-invariant features. In Proceedings of the 7th IEEE International Conference on Computer Vision 1150–1157.
58. Cortes, C. & V. Vapnik. 1995. Support-vector networks. *Mach. Learn.* **20:** 273–297.
59. Rosch, E., C. Simpson & R.S. Miller. 1976. Structural bases of typicality effects. *J. Exp. Psychol. Hum. Percept. Perform.* **2:** 491–502.
60. Attneave, F. 1950. Dimensions of similarity. *Am. J. Psychol.* **63:** 516–556.
61. Shepard, R.N. 1987. Towards a universal law of generalization for psychological science. *Science* **237:** 1317–1323.
62. Mur, M. *et al.* 2013. Human object-similarity judgments reflect and transcend the primate-IT object representation. *Front. Psychol.* **4:** 128.
63. Shepard, R.N. 1980. Multidimensional scaling, tree-fitting, and clustering. *Science* **210:** 390–398.
64. Peterson, J.C., J.T. Abbott & T.L. Griffiths. 2018. Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cogn. Sci.* **42:** 2648–2669.
65. Tversky, A. 1977. Features of similarity. *Psychol. Rev.* **84:** 327–352.
66. Shepard, R.N. & P. Arabie. 1979. Additive clustering: representation of similarities as combinations of discrete overlapping properties. *Psychol. Rev.* **86:** 87–123.
67. Navarro, D.J. & M.D. Lee. 2004. Common and distinctive features in stimulus similarity: a modified version of the contrast model. *Psychon. Bull. Rev.* **11:** 961–974.
68. Attarian, M., B.D. Roads & M.C. Mozer. 2020. Transforming neural network visual representations to predict human judgments of similarity. *arXiv preprint arXiv:2010.06512*.
69. Roads, B.D. & M.C. Mozer. 2021. Predicting the ease of human category learning using radial basis function networks. *Neural Comput.* **33:** 376–397.
70. Luce, R.D. 1959. *Individual Choice Behavior*. New York: John Wiley.
71. Jha, A., J. Peterson & T.L. Griffiths. 2020. Extracting low-dimensional psychological representations from convolutional neural networks. *arXiv preprint arXiv:2005. 14363*.
72. Saxe, A., J. McClelland & S. Ganguli. 2013. Learning hierarchical category structure in deep neural networks. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*.
73. Sanders, C.A. & R.M. Nosofsky. 2018. Using deep learning representations of complex natural stimuli as input to psychological models of classification. In *Proceedings of the 2018 Conference of the Cognitive Science Society, Madison*.
74. Bechberger, L. & K.-U. Kühnberger. 2019. Generalizing psychological similarity spaces to unseen stimuli. *arXiv preprint arXiv:1908.09260*.
75. Sanders, C.A. & R.M. Nosofsky. 2020. Training deep networks to construct a psychological feature space for a natural-object category domain. *Comput. Brain Behav.* **3:** 229–251.
76. Nosofsky, R.M., C.A. Sanders, B.J. Meagher & B.J. Douglas. 2018. Toward the development of a feature-space representation for a complex natural category domain. *Behav. Res. Methods* **50:** 530–556.
77. Peterson, J.C., P. Soulos, A. Nematzadeh & T.L. Griffiths. 2018. Learning hierarchical visual representations in deep neural networks using hierarchical linguistic labels. *arXiv preprint arXiv:1805.07647*.
78. Rosch, E. 1978. *Cognition and Categorization*. Hillsdale, NJ: Erlbaum.
79. Rosch, E. 1973. *Cognitive Development and Acquisition of Language*. Elsevier.

80. Wang, P. & G.W. Cottrell. 2015. Basic level categorization facilitates visual object recognition. *arXiv preprint arXiv:1511.04103*.
81. Rosenfeld, A., M.D. Solbach & J.K. Tsotsos. 2018. Totally looks like-how humans compare, compared to machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* 1961–1964.
82. Rosenfeld, A., R. Zemel & J.K. Tsotsos. 2019. High-level perceptual similarity is enabled by learning diverse tasks. *arXiv preprint arXiv:1903.10920*.
83. Cohen, H. & C. Lefebvre. 2005. *Handbook of Categorization in Cognitive Science*. Elsevier Science.
84. Plato. 1968. *The Republic*. New York: Basic Books.
85. Aristotle. 1984. *Categories*. Princeton, NJ: Princeton University Press.
86. Hull, C.L. 1920. Quantitative aspects of evolution of concepts: an experimental study. *Psychol. Monogr.* **28:** i–86.
87. Bruner, J.S., J.J. Goodnow & G.A. Austin. 1956. *A Study of Thinking*. New York: Wiley.
88. Rosch, E. 1973. Natural categories. *Cogn. Psychol.* **4:** 328–350.
89. Rosch, E. & C.B. Mervis. 1975. Family resemblances: studies in the internal structure of categories. *Cogn. Psychol.* **7:** 573–605.
90. Wittgenstein, L. 2009. *Philosophical Investigations*. John Wiley & Sons.
91. Posner, M.I. & S.W. Keele. 1968. On the genesis of abstract ideas. *J. Exp. Psychol.* **77:** 353–363.
92. Reed, S.K. 1972. Pattern recognition and categorization. *Cogn. Psychol.* **3:** 393–407.
93. Medin, D.L. & M.M. Schaffer. 1978. Context theory of classification learning. *Psychol. Rev.* **85:** 207–238.
94. Nosofsky, R.M. 1986. Attention, similarity, and the identification–categorization relationship. *J. Exp. Psychol. Gen.* **115:** 39–61.
95. Whittlesea, B.W. 1987. Preservation of specific experiences in the representation of general knowledge. *J. Exp. Psychol. Learn. Mem. Cogn.* **13:** 3–17.
96. Nosofsky, R.M., S.E. Clark & H.J. Shin. 1989. Rules and exemplars in categorization, identification, and recognition. *J. Exp. Psychol. Learn. Mem. Cogn.* **15:** 282–304.
97. Nosofsky, R.M. 1984. Choice, similarity, and the context theory of classification. *J. Exp. Psychol. Learn. Mem. Cogn.* **10:** 104–114.
98. Fried, L.S. & K.J. Holyoak. 1984. Induction of category distributions: a framework for classification learning. *J. Exp. Psychol. Learn. Mem. Cogn.* **10:** 234–257.
99. Aha, D.W. & R.L. Goldstone. 1992. Concept learning and flexible weighting. In Proceedings of the 14th Annual Conference of the Cognitive Science Society.
100. McKinley, S.C. & R.M. Nosofsky. 1995. Investigations of exemplar and decision bound models in large, ill-defined category structures. *J. Exp. Psychol. Hum. Percept. Perform.* **21**: 128–148.
101. Nosofsky, R.M. 1988. Similarity, frequency, and category representations. *J. Exp. Psychol. Learn. Mem. Cogn.* **14:** 54–65.
102. Jüttner, M. & I. Rentschler. 2000. Scale-invariant superiority of foveal vision in perceptual categorization. *Eur. J. Neurosci.* **12:** 353–359.
103. Palmeri, T.J. & R.M. Nosofsky. 2001. Central tendencies, extreme points, and prototype enhancement effects in ill-defined perceptual categorization. *Quart. J. Exp. Psychol. Sect. A* **54:** 197–235.
104. Stewart, N. & N. Chater. 2002. The effect of category variability in perceptual categorization. *J. Exp. Psychol. Learn. Mem. Cogn.* **28:** 893–907.
105. Maddox, W.T. & F.G. Ashby. 1993. Comparing decision bound and exemplar models of categorization. *Attent. Percept. Psychophys.* **53:** 49–70.
106. Rosseel, Y. 2002. Mixture models of categorization. *J. Math. Psychol.* **46:** 178–210.
107. Vanpaemel, W., G. Storms & B. Ons. 2005. A varying abstraction model for categorization. In Proceedings of the Annual Conference of the Cognitive Science Society 2277–2282.
108. Krizhevsky, A. & G. Hinton. 2009. Learning multiple layers of features from tiny images. Technical report. University of Toronto.
109. Annis, J., I. Gauthier & T.J. Palmeri. 2020. Combining convolutional neural networks and cognitive models to predict novel object recognition in humans. *J. Exp. Psychol. Learn. Mem. Cogn*. https://doi.org/10.1037/xlm0000968.
110. Gauthier, I. & M.J. Tarr. 1997. Becoming a "Greeble" expert: exploring mechanisms for face recognition. *Vis. Res.* **37:** 1673–1682.
111. Richler, J.J., J.B. Wilmer & I. Gauthier. 2017. General object recognition is specific: evidence from novel and familiar objects. *Cognition* **166:** 42–55.
112. Wong, A.C.-N., T.J. Palmeri, B.P. Rogers, *et al.* 2009. Beyond shape: how you learn about objects affects how they are represented in visual cortex. *PLoS One* **4:** e8405.
113. Richler, J.J. *et al.* 2019. Individual differences in object recognition. *Psychol. Rev.* **126:** 226–251.
114. Holmes, W.R., P. O'Daniels & J.S. Trueblood. 2020. A joint deep neural network and evidence accumulation modeling approach to human decision-making with naturalistic images. *Comput. Brain Behav.* **3:** 1–12.
115. Battleday, R.M., J.C. Peterson & T.L. Griffiths. 2017. Modeling human categorization of natural images using deep feature representations. *arXiv preprint arXiv:1711.04855*.
116. Guest, O. & B.C. Love. 2019. Levels of representation in a deep learning model of categorization. https://doi.org/10.1101/626374.
117. Schyns, P.G., R.L. Goldstone & J. Thibaut. 1998. Development of features in object concepts. *Behav. Brain Sci.* **21:** 1–54.
118. Recht, B., R. Roelofs, L. Schmidt & V. Shankar. 2018. Do CIFAR-10 classifiers generalize to CIFAR-10? *arXiv preprint arXiv:1806.00451*.
119. Kurakin, A., I. Goodfellow & S. Bengio. 2016. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*.
120. Szegedy, C. *et al.* 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

121. Goodfellow, I.J., J. Shlens & C. Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.
122. Nguyen, A., J. Yosinski & J. Clune. 2015. Deep neural networks are easily fooled: high confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 427–436.
123. Yamauchi, T. & A.B. Markman. 1998. Category learning by inference and classification. *J. Mem. Lang.* **39:** 124–148.
124. Yamauchi, T., B.C. Love & A.B. Markman. 2002. Learning nonlinearly separable categories by inference and classification. *J. Exp. Psychol. Learn. Mem. Cogn.* **28:** 585–593.
125. Peterson, J.C., R.M. Battleday, T.L. Griffiths & O. Russakovsky. 2019. Human uncertainty makes classification more robust. In *Proceedings of the IEEE International Conference on Computer Vision* 9617–9626.
126. Zhang, H., M. Cisse, Y.N. Dauphin & D. Lopez-Paz. 2017. mixup: beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.
127. Lake, B.M., R. Salakhutdinov, J. Gross & J. Tenenbaum. 2011. One shot learning of simple visual concepts. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.
128. Vinyals, O., C. Blundell, T. Lillicrap, *et al.* 2016. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems* 3630–3638.
129. Snell, J., K. Swersky & R. Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems* 4077–4087.
130. Welinder, P. *et al.* 2010. Caltech-UCSD birds 200.
131. Scott, T.R., K. Ridgeway & M.C. Mozer. 2019. Stochastic prototype embeddings. *arXiv preprint arXiv:1909.11702*.
132. Oh, S.J. *et al.* 2018. Modeling uncertainty with hedged instance embedding. *arXiv preprint arXiv:1810.00319*.
133. Allen, K.R., E. Shelhamer, H. Shin & J.B. Tenenbaum. 2019. Infinite mixture prototypes for few-shot learning. *arXiv preprint arXiv:1902.04552*.
134. Ren, M. *et al.* 2018. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*.
135. Lake, B.M., T.D. Ullman, J.B. Tenenbaum & S.J. Gershman. 2017. Building machines that learn and think like people. *Behav. Brain Sci.* **40:** e253.
136. Singh, P., J.C. Peterson, R.M. Battleday & T.L. Griffiths. 2020. End-to-end deep prototype and exemplar models for predicting human behavior. *arXiv preprint arXiv:2007.08723*.
137. Griffiths, T.L. 2015. Manifesto for a new (computational) cognitive revolution. *Cognition* **135:** 21–23.
138. Maaten, L.V.D. & G. Hinton. 2008. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9:** 2579–2605.