End-to-end Deep Prototype and Exemplar Models for Predicting Human Behavior

Pulkit Singh[1], Joshua C. Peterson[1], Ruairidh M. Battleday[1], Thomas L. Griffiths[1,2]
[1]Department of Computer Science, Princeton University
[2]Department of Psychology, Princeton University
{pulkit,joshuacp,battleday,tomg}@princeton.edu

Abstract

Traditional models of category learning in psychology focus on representation at the category level as opposed to the stimulus level, even though the two are likely to interact. The stimulus representations employed in such models are either hand-designed by the experimenter, inferred circuitously from human judgments, or borrowed from pretrained deep neural networks that are themselves competing models of category learning. In this work, we extend classic prototype and exemplar models to learn both stimulus and category representations jointly from raw input. This new class of models can be parameterized by deep neural networks (DNN) and trained end-to-end. Following their namesakes, we refer to them as Deep Prototype Models, Deep Exemplar Models, and Deep Gaussian Mixture Models. Compared to typical DNNs, we find that their cognitively inspired counterparts both provide better intrinsic fit to human behavior and improve ground-truth classification.

Keywords: category learning; deep neural networks; prototype models; exemplar models; Gaussian mixture models

Introduction

Categorization is central to cognition, and so models of category learning are ubiquitous in cognitive science (Kruschke, 1992; Nosofsky, 1984; Posner & Keele, 1968; Reed, 1972). Most models of category learning can be understood as methods for inferring the structure of different categories (*i.e.,* their representations). For example, prototype models represent a category as an idealization or abstraction—typically the mean—of its members (Reed, 1972), whereas exemplar models represent a category by all previously encountered members (Nosofsky, 1984). These category representations are by definition a function of the underlying stimulus representation, which is typically fixed in advance.

How the modeler should choose the "right" stimulus representation for any particular task is not clear. When the representations are designed by hand, they may be considerably biased. A popular alternative is to infer mental representations directly from judgments via multi-dimensional scaling (MDS; Shepard, 1980), but this process is sensitive to context (Medin, Goldstone, & Gentner, 1993), and when applied to downstream modeling, is limited to predicting behavior *from* behavior. Further, neither of these options provide a stimulus-to-representation mapping, nor a theory of how humans might acquire one. More problematic, however, is the observation that stimulus representations are unlikely fixed at all, and their structure may be influenced by the process of category learning itself (Schyns, Goldstone, & Thibaut, 1998).

More recent work (*e.g.,* Battleday, Peterson, & Griffiths, 2019; Lake, Zaremba, Fergus, & Gureckis, 2015; Peterson, Abbott, & Griffiths, 2018; Sanders & Nosofsky, 2018) leverages representations learned from deep neural networks to support psychological models. Unlike hand design or MDS, these networks both model learning and provide a mapping from raw inputs to low-, mid-, and high-level representations (LeCun, Bengio, & Hinton, 2015). However, these networks are typically trained for the task of categorization in a way that is not motivated by psychological theories. In particular, they are trained to discriminate categories via linear boundaries over their learned feature spaces, and do not explicitly represent the categories themselves. When we use them in service of psychological modeling, we are borrowing from an (implicit) theory of category learning that we do not espouse in order to support a different theory that we do.

Theories of feature learning in psychology do exist (*e.g.,* Austerweil & Griffiths, 2013; Rumelhart & Zipser, 1985), but they do not model the interaction between feature learning and category learning, and are often not computationally tractable for large, high-dimensional stimulus sets. While a variety of burgeoning unsupervised deep neural networks (*e.g.,* Donahue, Krähenbühl, & Darrell, 2016; Dumoulin et al., 2016; Ji, Henriques, & Vedaldi, 2019) provide competitive alternate candidates for stimulus representation that could support effective categorization, they are likewise fixed after training.

To address these shortcomings, we propose a probabilistic framework for incorporating feature learning into classic accounts of category learning, including prototype models, exemplar models, and a family of mixture models that define a continuum between these two extremes. Our framework can be instantiated via deep neural networks, backpropagating through each categorization model (*i.e.,* category representation learner) to a feature learning network that is learned simultaneously. We call these networks Deep Prototype Models (DPM), Deep Exemplar Models (DEM), and Deep Gaussian Mixture Models (DGMM). When trained on a popular benchmark dataset from machine learning, we find that our new family of models provide excellent out-of-sample fit to human uncertainty behavior for a large dataset of over $500,000$ judgments, and boost ground-truth classification accuracy compared to traditional deep neural networks.

Formalization of Category Learning

The problem of category learning can be formalized as inferring the probability that a stimulus, x, belongs to each of a set of C categories, $p(c_i \in C | x)$ (Anderson, 1991; Ashby & Alfonso-Reese, 1995). Bayes' rule implies that

$$p(c_i | x) = \frac{p(x|c_i)p(c_i)}{\sum_j p(x|c_j)p(c_j)}. \tag{1}$$

The prior over categories, $p(c_j)$, can reflect the frequency or saliency of each category. When it is uniform, the posterior is proportional to the likelihood of the stimulus given the category:

$$p(c_i | x) \propto p(x|c_i). \tag{2}$$

Our choice of a functional form for the likelihood, $p(x|c_i)$, specifies the strategy for representing categories, with each of the classical categorization models listed above corresponding to a particular functional form (Ashby & Alfonso-Reese, 1995).

A simple first choice for $p(x|c_i)$ is a Gaussian density, such that stimuli in each category are normally distributed with mean μ_i and covariance Σ_i. This implies that

$$p(c_i | x) \propto e^{-d(x, \mu_i, \Sigma_i)^2}, \tag{3}$$

where d is the Mahalanobis distance between the stimulus and the mean of the category given its covariance structure Σ_i. Gaussian likelihoods correspond to the probabilistic formulation of prototype models over integral psychological spaces, with category prototypes given by μ_i, and the similarity of a stimulus to the prototype inversely proportional to its distance. When $\Sigma_i \neq \mathbf{I}$, new variants of more complex prototype models emerge, including cases that are equivalent to models with curved decision boundaries (Ashby & Maddox, 1993; Battleday et al., 2019; Minda & Smith, 2001).

A much more flexible form for the likelihood is the kernel density estimator, which allows both the distribution of stimuli in each category to be arbitrarily distributed and the boundary between categories to be arbitrarily shaped. In this model, the probability distribution representing a category is composed of the normalized sum of a set of kernel distributions over the stimulus space. Taking a Gaussian kernel, we have that

$$p(c_i | x) \propto \sum_{x'} e^{-d(x, x', \mathbf{I})^2}, \tag{4}$$

where x' are kernel positions for c_i, and Σ_i is typically taken to be the identity matrix, $\mathbf{I}$; this reduces the Mahalanobis distance to the Euclidean distance. When all previously encountered members of category i are used as kernel locations, the Gaussian kernel density estimator is equivalent to the exemplar model (Ashby & Alfonso-Reese, 1995). Intuitively, this means that the probability that stimulus x comes from category c_i is inversely proportional to the sum of its distances to each member of that category. Other variations of exemplar

models may include alternate distance metrics, scaling parameters for the distance metric, and convex attentional transformations of x (Nosofsky, 1984).

Finally, we can consider a class of models that interpolate between these two extremes—of a single Gaussian likelihood for each category and one Gaussian kernel for each exemplar—by allowing the number of kernels per category, K, to vary between 1 and the number of exemplars in that category. This corresponds to a mixture of Gaussians, and allows the modeler to trade off between model accuracy and complexity given data (Griffiths, Sanborn, Canini, & Navarro, 2012). Although less commonly used, Gaussian mixture models have also been proposed as models of human category learning (Rosseel, 2002), and are special cases of other influential categorization frameworks (Anderson, 1991; Griffiths, Canini, Sanborn, & Navarro, 2007).

Incorporating Feature Learning

Incorporating feature learning into classic models implies that, in addition to a learnable category representation, a joint model must also specify a learnable transformation of stimulus inputs, $\phi(\cdot)$, over which the categorization model will operate. The idea of inferring representations of stimuli for input into categorization models has been employed for some time via strategies such as MDS (Shepard, 1980), or by searching over a range of feature extraction algorithms (Battleday et al., 2019). However, although these methods yield a transformed input representation, $\phi(x)$, they do not provide the corresponding transformation $\phi(\cdot)$ itself, nor a means to learn it. The transformation itself is important because it both explains part of human learning and allows a model to generalize to new stimuli.

For complex naturalistic stimuli, we want the flexibility to learn a transformation $\phi(\cdot)$—and thus, a resulting representation $\phi(x)$—that can best support a particular categorization strategy or category representation. However, if $\phi(\cdot)$ is learned incrementally, then each individual evaluation of the transformation will require either re-estimation of category means or mixture centers, or re-transformation of all exemplars in the dataset, both of which seem psychologically implausible and computationally intractable. Alternatively, we can parameterize our likelihood function for each category i by a set of K learnable centers $z_i = (z_1, ..., z_K)$ for category i, and take the likelihood function $p(\phi(x)|c_i, z_i))$ to be the Gaussian kernel density estimator proposed above, with kernels at locations z_i for category i. The task of category learning can therefore be re-framed as inferring the probability

$$\arg\max_{\phi, z} p(c_i | \phi(x)) \propto p(\phi(x)|c_i, z_i)), \tag{5}$$

When $K = 1$, the number of centers per category corresponds to learned prototypes for each category. Assuming an equal number of stimuli per class, when $K = \frac{N}{C}$ there are as many centers as stimuli per class, corresponding to learned exemplars. For C total classes, when $\frac{N}{C} > K > C$, there are an arbitrary number of centers, and while they could be thought of as
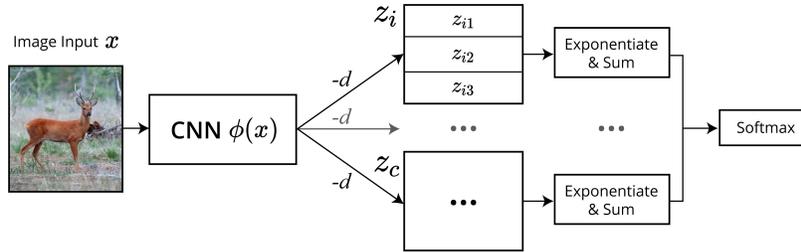
Figure 1: General schematic for deep categorization models.

subsets of exemplars, it may be most appropriate to think of them as the centers of a Gaussian mixture density. In general, we are now interested in the probability

$$p(c_i \,|\, \phi(x), z_i) \propto \sum_{z' \in z_i} -d(\phi(x), z', \mathbf{I})^2. \qquad (6)$$

Note that exponentiation and normalization are omitted for clarity, since the proportional relationship still holds. When $K = 1$, the sum contains only a single term, corresponding to a Gaussian density (prototype model), when $\frac{N}{C} > K > C$, the sum corresponds to a Gaussian mixture density with uniform weights, and when $K = \frac{N}{C}$, the sum corresponds to a Gaussian-kernel density (exemplar model).

Deep Categorization Models

While transformation $\phi(\cdot)$ and centers z_i could be estimated through a variety of methods, we choose to express both using a deep convolutional neural network (CNN; Krizhevsky, Sutskever, & Hinton, 2012) with parameter matrices (θ_1, θ_2), which among other reasons are highly scalable and efficient for featurizing image inputs. The transformation ϕ_{θ_1} is instantiated as a convolutional neural network and Z_{θ_2} parameterizes a final classification layer via learned centers z_i for each category. A visualization of our deep categorization models is shown in Figure 1. Particular variants of the model are explained in detail below.

Deep Prototype Models. In the case of Gaussian likelihoods, the covariance matrix Σ_j determines whether the distance function is Euclidean or Mahalanobis. This difference corresponds to constraints on the possible shape of the Gaussian category densities that will be estimated. When Σ_j is the identity $\mathbf{I}$ (denoted as $\Sigma_{\mathbf{I}}$ hereafter), we assume that category distributions are isometric Gaussians (*i.e., d*-dimensional spheres in the stimulus or feature space). When the diagonal of Σ_j is a scalar multiple of $\mathbf{I}$ (*i.e.,* the diagonal is filled by the same constant that does not need to be 1), denoted as Σ_C hereafter, the size of the sphere and variance of the category is allowed to vary across categories. Finally, when the elements of the diagonal of Σ_j can vary freely, denoted as Σ_A hereafter, the category distributions are allowed to stretch into d-dimensional ellipsoids that are aligned with the axes of the stimulus space. An additional case is also possible, wherein Σ_j is a full covariance matrix, and dimensions

of the Gaussian are allowed to be correlated. However, we leave this much more expressive variant to future work.

Notably, DPMs are formally similar to a training objective called *center loss* (Wen, Zhang, Li, & Qiao, 2016), although the latter re-estimates class means for each mini-batches during training. They are also related to prototypical networks (Scott, Ridgeway, & Mozer, 2019; Snell, Swersky, & Zemel, 2017), which learn representations such that averages of exemplars from unseen classes support few-shot learning.

Deep Gaussian Mixture Models. DGMMs were implemented as multi-center versions of the DPMs described above, but distances are summed across all centers for each class, and the distance function is always Euclidean. This model contrasts with related ones that are instead trained asynchronously (Variani, McDermott, & Heigold, 2015), encourage large margins (Wan, Zhong, Li, & Chen, 2018), or focus on multiple layers of latent variables (Van den Oord & Schrauwen, 2014; Viroli & McLachlan, 2019).

Deep Exemplar Models. DEMs were implemented as DG-MMs with as many centers as training datapoints. Exemplars were initialized as the initial random projections (a result of random starting weights) of the training data through the untrained neural network. It should be noted that another popular neural network model of category learning (which also does not incorporate feature learning) known as AL-COVE (Kruschke, 1992) is a special case of our models where centers are arranged in a grid or randomly initialized in a fixed stimulus space, but not trained. DEMs are also related to deep Gaussian process hybrid deep networks (Bradshaw, Matthews, & Ghahramani, 2017), which are fitted with stochastic variational inference.

Methods

Dataset. While there are a variety of large natural image datasets from computer vision that may be appropriate for training, (*e.g.,* Barbu et al., 2019; Deng et al., 2009), we focus on CIFAR-10 (Krizhevsky & Hinton, 2009), a popular benchmark; albeit one with relatively small images in terms of pixel size. We choose CIFAR-10 both because it allows us to train many hundreds of models and replications to convergence in a reasonable amount of time (about 5-7 hours each),

	Baseline		Prototype (DPM)						DGMM		Exemplar (DEM)	
	ResNet	*AllCNN*	ResNet			*AllCNN*			*Resnet*	*AllCNN*	*ResNet*	*AllCNN*
			$\Sigma_{\mathbf{I}}$	Σ_C	Σ_A	$\Sigma_{\mathbf{I}}$	Σ_C	Σ_A				
`CIFAR-10` (Accuracy)	90.3	92.3	90.3	90.3	90.4	92.3	92.3	91.8	90.4	**92.5**	90.2	92.1
`CIFAR-10H` (Error)	.88	.78	.72	.77	.84	.55	.63	.67	.72	**.45**	.74	.57

Table 1: Mean `CIFAR-10` validation accuracy and `CIFAR-10H` error (crossentropy; fit to human uncertainty) for each model.

	Baseline		Prototype (DPM)						DGMM		Exemplar (DEM)	
	ResNet	*AllCNN*	ResNet			*AllCNN*			*Resnet*	*AllCNN*	*ResNet*	*AllCNN*
			$\Sigma_{\mathbf{I}}$	Σ_C	Σ_A	$\Sigma_{\mathbf{I}}$	Σ_C	Σ_A				
`CIFAR-10` (Accuracy)	90.3	92.3	90.4	90.5	90.5	92.4	92.4	91.9	90.8	**92.8**	90.6	92.5
`CIFAR-10H` (Error)	.88	.76	.72	.77	.84	.53	.63	.67	.71	**.43**	.72	.48

Table 2: Best `CIFAR-10` validation accuracy and `CIFAR-10H` error (crossentropy; fit to human uncertainty) for each model.

and because it allows for a rigorous comparison to an existing, large human behavioral dataset (see **Evaluation** below). `CIFAR-10` contains 5000 training and 1000 validation images for each of ten categories: *airplane*, *automobile*, *bird*, *cat*, *deer*, *dog*, *frog*, *horse*, *ship*, and *truck*.

CNN Architectures. To learn transformation ϕ_{θ_1} over input images, we employ two architectures: ResNet (He, Zhang, Ren, & Sun, 2016) and All-CNN (Springenberg, Dosovitskiy, Brox, & Riedmiller, 2014). All-CNN is a surprisingly parsimonious (*i.e.,* uses only convolutional layers), low-parameter, stable, high-accuracy network that can be trained quickly. Our only architectural modification is the addition of batch normalization layers after each rectified linear (ReLU) activation function on top of each convolutional layer, which significantly reduced training time. Including the new layers before the ReLUs was less effective in our tests. For ResNet, we use the 20-layer, *v2* version, which is by no means the best of its cousins, but has a high size-to-accuracy ratio. The final representation layers in All-CNN and ResNet are 192- and 256-dimensional respectively. We also use each unmodified architecture as a baseline model, with the original fully-connected classification layer at the end.

For both networks, we do not mirror the original training schemes exactly, which involve different optimizers and complicated learning rate schedules that would interact with our new architectural changes; instead, we hold them constant to reduce the search space. In particular we used a simple gradient descent optimizer with a learning rate of 0.01, decay value of 1e-6, and Nesterov momentum with a value of 0.9. These values were chosen before any baselines or novel models were trained.

Training. We train each of our models in the typical supervised manner. While categorization models in psychology are typically fit directly to human behavior, we follow the machine learning paradigm of presenting only single labels for each image. This mimics a type of supervision that humans receive, since humans do not learn categories by fitting to their own behavior. As explained below, we evaluate

our models' fit to humans *after* learning, requiring that our models learn to behave like humans without being explicitly trained to do so.

Evaluation. Following standard practice in machine learning, we evaluate the accuracy of our models on a held out set of images and ground-truth labels (in our case, the CIFAR10 validation set described above). This is mostly to ensure that our baselines are valid and our models are competitive in this regard. However, we are mainly interested in fit to human behavior. Peterson, Battleday, Griffiths, and Russakovsky (2019) provide a useful dataset in this regard, which measures human uncertainty for each image of the 10,000-image `CIFAR-10` validation, which they call `CIFAR-10H`.

`CIFAR-10H` contains 10-way classification judgments from 50 subjects for each image (over 500,000 total judgments), yielding probability distributions that reflect human uncertainty at the image level. For example, for a particular image, which are small and often somewhat ambiguous, 60% of subjects may have chosen *dog*, and the other 40% may have chosen *cat*. These proportions are directly comparable to the output of a network when expressed as probabilities. The authors validate this dataset by showing that training directly on it (as opposed to typical "one-hot" vectors) generalizes to human uncertainty for held out images, generalizes in terms of accuracy to a series of large and challenging held out validation datasets, and confers robustness to adversarial attacks. This both defines a crucial natural-image benchmark for predicting human behavior, and also introduces a new metric for objectively evaluating image classification models. To apply this dataset as evaluation, crossentropy (error) between the probability distributions output by a trained model y_i and the human probabilities h_i for each image in the validation set is computed: $-\sum_{i=1}^{C} h_i \log(y_i)$.

Results

Performance

For all results, we report both the best (Table 1) and the averages (Table 2) of 10 training runs, which exhibit the same pattern of results. We first note that all but one of our eight
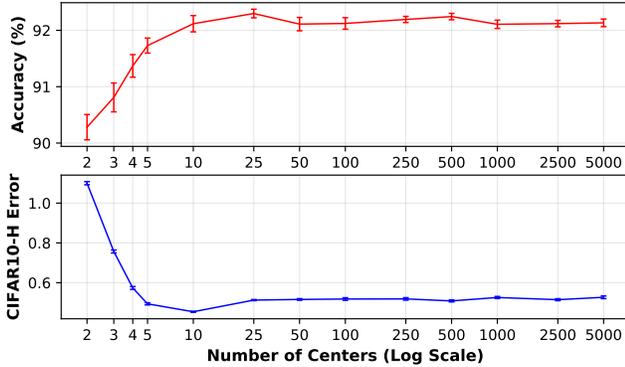
Figure 2: Accuracy and human fit as no. model centers varies.

models outperform the standard convolutional neural network baseline for accuracy on `CIFAR-10` ground-truth labels, with the Deep Gaussian Mixture Model scoring highest. As accuracy gains not trivial to obtain for well-studied architectures, these are encouraging results.

However, as we are primarily interested in providing high-quality models of human categorization for more complex stimuli, the main criterion of model success is the fit to human-derived labels for these images from `CIFAR-10H`. Here we also find that all of our models outperform the neural network baselines, but with a much greater improvement, particularly for the DGMMs and DEMs. For comparison, in the study that introduced this `CIFAR10-H`, a ResNet architecture pretrained on `CIFAR10` and further fine-tuned on subsets of the `CIFAR10-H` human label distributions achieved a cross entropy loss of 0.35 (Peterson et al., 2019). Our models achieve nearly the same loss with none of the extra supervision, in most cases with lower accuracy (the authors do extensive hyperparameter search), and without any additional fine-tuning of hyper-parameters.

Optimal Number of Centers

In the previous section, we make a distinction between models with single clusters for each category (prototypes), models with as many clusters as training datapoints (exemplars), and anything in between (mixture centers). For DGMMs, we reported only the best models. To explore the optimal number of centers for predicting human uncertainty (and obtaining highest accuracy) more finely, we vary the number of centers from 2 to 5000 for All-CNN and plot them against each performance metric in Figure 2, averaged over 10 runs. We first note that both accuracy and fit to humans are a smooth function of the number of centers and exhibit a clear maximum. To maximize accuracy, roughly 25 centers are optimal; to maximize fit to human uncertainty, less than half of that are needed (roughly 10). This suggests that both can be maximized given some number of centers $10 < M < 25$, and provides clearer evidence that GMM models of category learning, a middle ground between infamously opposed prototype and exemplar models, may best model joint human feature and category representations for natural images.
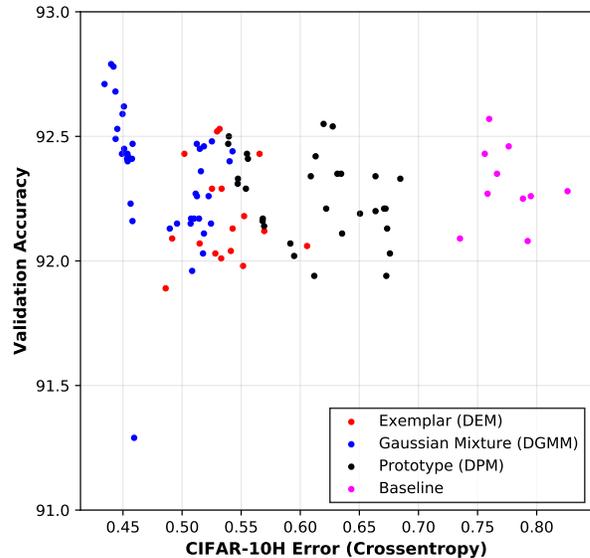

Figure 3: Validation accuracy × fit to human uncertainty.

Accuracy Versus Fit to Humans

Compared to baseline, our models are both more accurate and a much better fit to human uncertainty. While our models contain more parameters, the complexity (*i.e.,* parameter count) of each model does not appear to be related to performance: the most complex DPMs (*i.e.*, with axis-aligned covariance matrices) and the most complex multi-center models (*i.e.,* DEMs) were not the best models. The optimal number of centers was also quite modest as we saw in the previous section. Further, accuracy does not appear to be related to human fit at all. In Figure 3, we plot accuracy against error on `CIFAR10-H` for all 90 All-CNN models that we trained and find essentially no discernible relationship. Instead, fit to humans appears to be solely a function of categorization model (identified by color).

Visualizing Category Representations

Lastly, we visualize representations and centers using t-SNE for two illustrative model classes: identity-covariance DPMs, with a single center per class, and a DGMM with 100 centers per class. For the DPMs we expect to see a single point (dark color) representing the prototype for each category near the center of its members. For DGMMs, there are two possibilities: (1) a degenerate case where the model learned many near duplicates of the mean of the category, or (2) a more interesting case where centers have an arbitrary nontrivial distribution that supports a nonlinear classification boundary. The results are shown in Figure 4. DPMs appear largely unimodal, as expected, and DGMMs indeed exhibit an interesting and non-degenerate pattern: centers are reasonably spread out, sometimes in nonlinear shapes or multimodal clusters, and not always located at the center of class members. The distribution of category members is also multimodal: subsets of centers are sometimes located distantly from the others in order to accommodate these flexible distributions.

Discussion

Traditional models of category learning in psychology are evaluated in the context of underlying stimulus representations that are chosen independently, which fails to model their interaction, and may significantly bias model comparison. We directly address this issue by learning the most appropriate representation for each categorization strategy.

Our approach appears to have been successful in that DCMs can be trained in a stable manner with no apparent drawbacks. Even networks that must employ up to 5000 centers per category had no trouble learning useful stimulus representations in parallel. Moreover, nearly all of our models exhibit patterns of uncertainty in their post-training predictions that match people, and most of them improve accuracy, which is no easy feat given a fixed, well-known architecture in computer vision. Since we expected the former results but not the latter, we did not aim to augment the absolute state-of-the-art CNNs, and so cannot yet fully assess the utility of our models on that tier. However, the results we obtained were remarkably consistent across our ten training runs. The pattern of results reported in this paper were obtained in one large training batch, and were not revised to find hyperparameters more advantageous to our models' performance.

While the accuracy gains obtained by our models were modest, improvements in fit to human uncertainty were relatively large compared to previous work (Peterson et al., 2019) that required extensive human-based supervision to obtain comparable results. This result is important because agreement between human and model uncertainty patterns is both a better predictor of robustness and generalization than accuracy (Peterson et al., 2019), and a crucial verification of psychological theories of category learning that posit explicit category representation. Although producing datasets like CIFAR10-H is not as straightforward for datasets of larger images with less ambiguity (and therefore less human uncertainty), the success of our category learning models on this dataset provides way to potentially improve network performance on any dataset, since fitting directly to human behavior is not necessary.

Conclusion

In the current work, we have proposed a set of models that generalize several seminal accounts of categorization to include feature learning, making use of deep neural networks as our workhorses. One advantage of doing so is that it enables us to study more complex and naturalistic stimulus sets, while ensuring stimulus and category representations cohere well with inaccessible human psychological representations. We see such integrations as the next step in the synergistic relationship between modern machine learning and cognitive science. In particular, we see this successful experiment as motivation for cognitive modelers to take inspiration from scalable methods, and as evidence to machine learning practitioners that insights into human cognition can tangibly inform the design of AI.
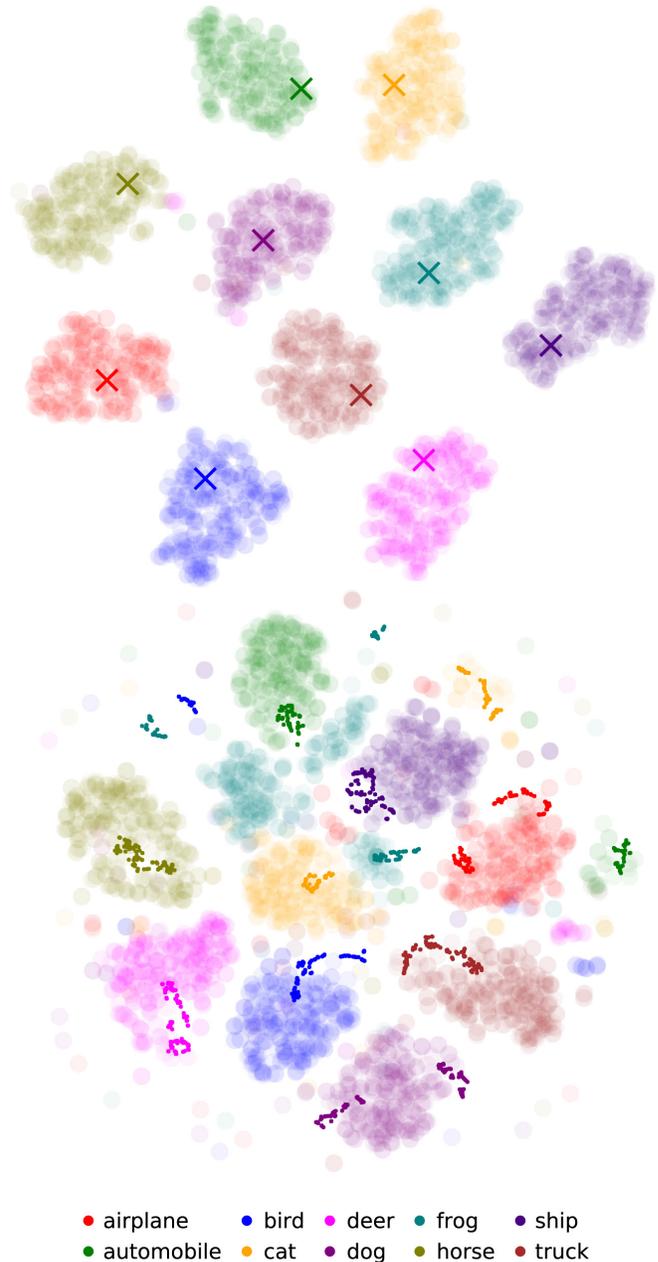

Figure 4: t-SNE embeddings of 3000 random stimuli and all per-category centers for a DPM (top) and DGMM (bottom).

Acknowledgements

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*, 216-233.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, *37*, 372-400.

Austerweil, J. L., & Griffiths, T. L. (2013). A nonparametric Bayesian framework for constructing flexible feature representations. *Psychological Review*, *120*(4), 817-851.

Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., ... Katz, B. (2019). Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems* (pp. 9448–9458).

Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2019). Capturing human categorization of natural images at scale by combining deep networks and cognitive models. *arXiv preprint arXiv:1904.12690*.

Bradshaw, J., Matthews, A. G. d. G., & Ghahramani, Z. (2017). Adversarial examples, uncertainty, and transfer testing robustness in gaussian process hybrid deep networks. *arXiv preprint arXiv:1707.02476*.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Donahue, J., Krähenbühl, P., & Darrell, T. (2016). Adversarial feature learning. *arXiv preprint arXiv:1605.09782*.

Dumoulin, V., Belghazi, I., Poole, B., Mastropietro, O., Lamb, A., Arjovsky, M., & Courville, A. (2016). Adversarially learned inference. *arXiv preprint arXiv:1606.00704*.

Griffiths, T. L., Canini, K., Sanborn, A., & Navarro, D. (2007). Unifying rational models of categorization via the hierarchical Dirichlet process. In *Proceedings of the Twenty-Ninth Annual Meeting of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

Griffiths, T. L., Sanborn, A. N., Canini, K. R., & Navarro, D. J. (2012, March 22). Categorization as nonparametric bayesian density estimation. In *The probabilistic mind.* United Kingdom: Oxford University Press. doi: 10.1093/acprof:oso/9780199216093.003.0014

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

Ji, X., Henriques, J. F., & Vedaldi, A. (2019). Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 9865–9874).

Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto.

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

Kruschke, J. K. (1992). ALCOVE: an exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22-44.

Lake, B., Zaremba, W., Fergus, R., & Gureckis, T. M. (2015). Deep neural networks predict category typicality ratings for images. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society.*

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

Medin, D., Goldstone, R., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, *100*, 254-278.

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: the effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(3), 775.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(1), 104.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive Science*, *42*(8), 2648–2669.

Peterson, J. C., Battleday, R. M., Griffiths, T. L., & Russakovsky, O. (2019, October). Human uncertainty makes classification more robust. In *The IEEE International Conference on Computer Vision.*

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*, 353-363.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*, 393-407.

Rosseel, Y. (2002). Mixture models of categorization. *Journal of Mathematical Psychology*, *46*, 178-210.

Rumelhart, D. E., & Zipser, D. (1985). Feature discovery by competitive learning. *Cognitive science*, *9*(1), 75–112.

Sanders, C. A., & Nosofsky, R. M. (2018). Using deep learning representations of complex natural stimuli as input to psychological models of classification. In *Proceedings of the Annual Meeting of the Cognitive Science Society.*

Schyns, P. G., Goldstone, R. L., & Thibaut, J. (1998). Development of features in object concepts. *Behavioral and Brain Sciences*, *21*, 1-54.

Scott, T. R., Ridgeway, K., & Mozer, M. C. (2019). Stochastic prototype embeddings. *arXiv preprint arXiv:1909.11702*.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, *210*, 390-398.

Shepard, R. N. (1987). Towards a universal law of generalization for psychological science. *Science*, *237*, 1317-1323.

Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems* (pp. 4077–4087).

Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*.

Van den Oord, A., & Schrauwen, B. (2014). Factoring variations in natural images with deep gaussian mixture models. In *Advances in Neural Information Processing Systems* (pp. 3518–3526).

Variani, E., McDermott, E., & Heigold, G. (2015). A gaussian mixture model layer jointly optimized with discriminative features within a deep neural network architecture. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4270–4274).

Viroli, C., & McLachlan, G. J. (2019). Deep gaussian mixture models. *Statistics and Computing*, *29*(1), 43–51.

Wan, W., Zhong, Y., Li, T., & Chen, J. (2018). Rethinking feature distribution for loss functions in image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 9117–9126).

Wen, Y., Zhang, K., Li, Z., & Qiao, Y. (2016). A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision* (pp. 499–515).