Capturing human category representations by sampling in deep feature spaces

Joshua C. Peterson[1] (jpeterson@berkeley.edu)
Jordan W. Suchow[1] (suchow@berkeley.edu)
Krisha Aghi[2] (kaghi@berkeley.edu)
Alexander Y. Ku[1] (alexku@berkeley.edu)
Thomas L. Griffiths[1] (tom_griffiths@berkeley.edu)
Department of Psychology[1], Helen Wills Neuroscience Institute[2]
University of California, Berkeley, CA 94720 USA

Abstract

Understanding how people represent categories is a core problem in cognitive science. Decades of research have yielded a variety of formal theories of categories, but validating them with naturalistic stimuli is difficult. The challenge is that human category representations cannot be directly observed and running informative experiments with naturalistic stimuli such as images requires a workable representation of these stimuli. Deep neural networks have recently been successful in solving a range of computer vision tasks and provide a way to compactly represent image features. Here, we introduce a method to estimate the structure of human categories that combines ideas from cognitive science and machine learning, blending human-based algorithms with state-of-the-art deep image generators. We provide qualitative and quantitative results as a proof-of-concept for the method's feasibility. Samples drawn from human distributions rival those from state-of-the-art generative models in quality and outperform alternative methods for estimating the structure of human categories.

Keywords: categorization; neural networks; Markov Chain Monte Carlo

Introduction

Categorization is a central problem in cognitive science and concerns why and how we divide the world into discrete units at various levels of abstraction. The biggest challenge for studying human categorization is that the content of mental category representations cannot be directly observed, which has led to development of laboratory methods for estimating this content from human behavior. Because these methods rely on small sets of artificial stimuli with handcrafted or low-dimensional feature sets, they are ill-suited to the study of categorization as an intelligent process, which is principally motivated by robust human categorization performance in complex ecological settings (Nosofsky et al., 2017).

One of the challenges of applying laboratory methods to realistic stimuli such as natural images is finding a way to represent them. Deep learning models, such as convolutional neural networks, discover features that can be used to represent complex images compactly and perform well on a range of computer vision tasks (LeCun et al., 2015). It may be possible to express human category structure using these features, an idea supported by recent work in cognitive science (Lake et al., 2015; Peterson et al., 2016).

Ideally, experimental methods could be combined with state-of-the-art deep learning models to estimate the structure of human categories with as few assumptions as possible, and while avoiding the problem of dataset bias. In what follows, we propose a method that uses a human in the loop to estimate arbitrary distributions over complex feature spaces, adapting an existing experimental paradigm to exploit advances in deep architectures to capture the precise structure of human category representations and iteratively sharpen them. Such knowledge is crucial to forming an ecological theory of intelligent categorization behavior and to providing a ground-truth benchmark to guide future work in machine learning.

Background

Deep neural networks for images Deep neural networks are modern instantiations of classic multilayer perceptrons, and represent a powerful class of machine learning model. DNNs can be trained efficiently through gradient descent and structurally specialized for particular domains (LeCun et al., 2015). In the image domain, deep convolutional neural networks (CNNs; LeCun et al., 1989) excel in classic computer vision tasks, including natural image classification (Krizhevsky et al., 2012). CNNs exploit knowledge of the input domain by learning a hierarchical set of translation-invariant image filters. The resulting representations, real-valued feature vectors, are surprisingly general and outperform other methods in explaining complex human behavior (Lake et al., 2015; Peterson et al., 2016).

Generative Adversarial Networks (GANs; Goodfellow et al., 2014) and Variational Autoencoders (VAEs; Kingma & Welling, 2013) provide a generative approach to modeling the content of natural images. Importantly, though the approaches differ considerably, each approach makes use of a network (called a "decoder" or "generator") that learns a deterministic function that maps samples from a known noise distribution $p(z)$ (e.g., a multivariate Gaussian) to samples from the true image distribution $p(x)$. This can be thought of as mapping a relatively low-dimensional feature representation z to a relatively high-dimensional image x. Sampling new images from these networks is as simple as passing Gaussian noise into the learned decoder. In addition, because of its simple form, the resulting latent space z tends to be easy to traverse meaningfully (i.e., an intrinsic linear manifold) and can be readily visualized via the decoder, a property we exploit presently.

Estimating the structure of human categories Methods for estimating human category templates have existed for some time. In psychophysics, the most popular and well-understood method is known as *classification images* (CI; Ahumada, 1996).

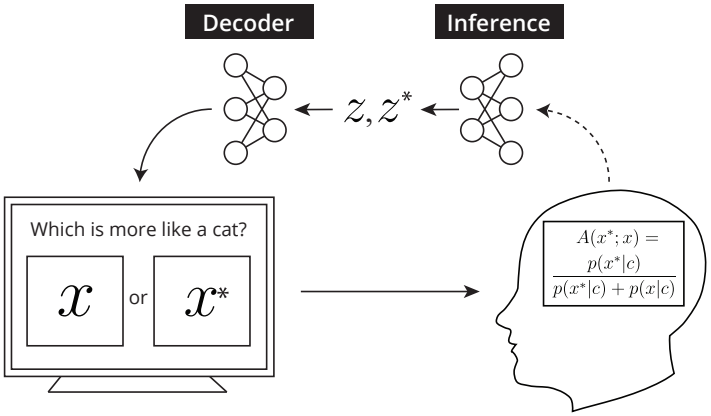

Figure 1: Deep MCMCP. A current state z and proposal z^* (top middle) are fed to a pretrained deep image generator/decoder network (top left). The corresponding decoded images x and x^* for the two states are presented to human raters on a computer screen (leftmost arrow and bottom left). Human raters then view the images in an experiment (bottom middle arrow) and act as part of an MCMC sampling loop, choosing between the two states/images in accordance with the Barker acceptance function (bottom right). The chosen image can then be sent to the inference network (rightmost arrow) and decoded in order to select the state for the next trial, however this step is unnecessary when we know exactly which states corresponds to which images.

In the classification images experimental procedure, a human participant is presented with images from two categories, A and B, each with white noise overlaid, and asked to select the stimulus that corresponds to the category in question. On most trials, the participant will obviously select the exemplar generated from the category in question. However, if the added white noise significantly perturbs features of the image that are important to making the distinction, they may fail. Exploiting this, we can estimate the decision boundary from a number of these trials using the simple formula:

$$(n_{AA} + n_{BA}) - (n_{AB} + n_{BB}), \tag{1}$$

where n_{XY} is the average of the noise across trials where the correct class is X and the observer chooses Y.

Vondrick et al. (2015) used a variation on classification images using deep image representations that could be inverted back to images using an external algorithm. In order to avoid dataset bias introduced by perturbing real class exemplars, white noise in the feature space was used to generate stimuli. In this special case, category templates reduce to $n_A - n_B$. On each trial of the experiment, participants were asked to select which of two images (inverted from feature noise) most resembled a particular category. Because the feature vectors for all trials were random, thousands of stimuli could be rendered in advance of the experiment using relatively slow methods that require access to large datasets. This early inversion method was applied to mean feature vectors for thousands of positive choices in the experiments and yielded qualitatively decipherable category template images, as well as better objective classification decision boundaries that were guided human bias. Under the assumption that human category distributions are Gaussian with equal variance, this method yields a vector that aligns with the nearest-mean decision boundary, although a massive number of human trials are required.

Markov Chain Monte Carlo with People (MCMCP; Sanborn & Griffiths, 2007), an alternative to classification images, is an experimental procedure in which humans act as a valid acceptance function A in the Metropolis–Hastings algorithm, exploiting the fact that Luce's choice axiom, a well-known model of human choice behavior, is equivalent to the Barker acceptance function (see equation in Figure 1). On the first trial, a stimulus x is drawn arbitrarily from the parameter space and compared to a new proposed stimulus x^* that is nearby in that parameter space. The participant makes a forced choice as to which is the better exemplar of some category (e.g., dog), acting as the acceptance function $A(x^*; x)$. If the initial stimulus is chosen, the Markov chain remains in that state. If the proposed stimulus is chosen, the chain moves to the proposed state. The process then repeats until the chain converges to the target category distribution $p(x|c)$. In practice, convergence is assessed heuristically, or limited by the number of human trials that can be practically obtained.

MCMCP has been successfully employed to capture a number of different mental categories (Sanborn & Griffiths, 2007; Martin et al., 2012), and though these spaces are higher-dimensional than those in previous laboratory experiments, they are still relatively small and artificial compared to real images. Unlike classification images, this method makes no assumptions about the structure of the category distributions and thus can estimate means, variances, and higher order moments. Therefore, we take it as a starting point for the current method.

MCMCP in deep feature spaces

The typical MCMCP experiment is effective so long as noise can be added to dimensions in the stimulus parameter space to create meaningful changes in content. In the case of natural images, noise in the space of all pixel intensities is very unlikely to modify the stimulus in meaningful ways. Instead, we propose perturbing images in a deep feature space that captures only essential variation. Since trials in an MCMCP experiment are not independent, we employ real-time, web-accessible generative adversarial networks to render high quality inversions from their latent features. The mapping from features to images learned by a GAN is deterministic, and therefore MCMCP in low-dimensional feature space approximates the same process in high-dimensional image space. The resulting judgments (samples) approximate distributions that both derive arbitrary human category boundaries for natural images and can be sampled from to create

images, yielding new human-like generative image models. A schematic of this procedure is illustrated in Figure 1.

There are several theoretical advantages to our method over previous efforts. First, MCMCP can capture arbitrary distributions, so it is not as sensitive to the structure of the underlying low-dimensional feature space and should provide better category boundaries than classification images when required. This is important when using various deep features spaces that were learned with different training objectives and architectures. MCMC inherently spends less time in low probability regions and should in theory waste fewer trials. Having generated the images online and as a function of the participant's decisions, there is no dataset or sampling bias, and auto-correlation can be addressed by removing temporally adjacent samples from the chain. Finally, using a deep generator provides drastically clearer samples than shallow reconstruction methods, and can be trained end-to-end with an inference network that allows us to categorize new images using the learned distribution.

Experiments

For our experiments, we explored two image generator networks trained on various datasets. Since even relatively low-dimensional deep image embeddings are large compared to controlled laboratory stimulus parameter spaces, we use a hybrid proposal distribution in which a Gaussian with a low variance is used with probability P and a Gaussian with a high variance is used with probability $1-P$. This allows participants to both refine and escape nearby modes, but is simple enough to avoid excessive experimental piloting that more advanced proposal methods often require.

Participants in all experiments completed exactly 64 trials (image comparisons), collectively taking about 5 minutes, containing segments of several chains for multiple categories. The order of the categories and chains within those categories were always interleaved. Each participant's set of chains for each category were initialized with the previous participants final states, resulting in large, multi-participant chains. All experiments were conducted on Amazon Mechanical Turk. If a single image did not load for a single trial, the data for the subject undergoing that trial was completely discarded, and a new subject was recruited to continue on from the original chain state.

Experiment 1: Initial test with face categories

Methods We first test our method using DCGAN (Radford et al., 2015) trained on the Asian Faces Dataset. We chose this dataset because it requires a deep architecture to produce reasonable samples (unlike MNIST, for example), yet it is constrained enough to test-drive our method using a relatively simple latent space. Four chains for each of four categories (male, female, happy, and sad) were used. Proposals were generated from an isometric Gaussian with a standard deviation of 0.25 50% of the time, and 2 otherwise. In addition, we conducted a baseline in which two new initial state proposals were drawn on every trial, and were independent of previous trials (classification images). The final dataset contained 50 participants and over $3,200$ trials (samples) in total for all chains. The baseline classification images (CI) dataset contained the same number of trials and participants.

Results MCMCP chains are visualized using Fisher Linear Discriminant Analysis in Figure 2, along with the resulting averages for each chain and each category. Chain means within a category show interesting variation, yet converge to similar regions in the latent space as expected. Figure 2 also shows visualizations of the mean faces for both methods in the final two columns. MCMCP means appear to have converged quickly, whereas CI means only moderately resemble their corresponding category (e.g., the MCMCP mean for "happy" is fully smiling, while the CI mean barely reveals teeth). All four CI means appear closer to a mean face, which is what one would expect from averages of noise. We validated this improvement with a human experiment in which 30 participants made forced choices between CI and MCMCP means. The results are reported in Figure 3. MCMCP means are consistently highly preferred as representations of each category as compared to CI. This remained true even when an additional 50 participants (total of 100) completed the CI task, obtaining twice as many image comparison trials as with MCMCP.

Experiment 2: Larger networks & larger spaces

The results of Experiment 1 show that reasonable category templates can be obtained using our method, yet the complexity of the stimulus space used does not rival that of large object classification networks. In Experiment 2, we tackled a more challenging (and interesting) form of the problem. To do this, we employed a bidirectional generative adversarial network (BiGAN; Donahue et al., 2016) trained on the 1.2 million-image ILSVRC12 dataset (64×64 center-cropped). BiGAN includes an inference network, which regularizes the rest of the model and produces unconditional samples competitive with the state-of-the-art. This also allows for the later possibility of comparing human distributions with other networks as well as assessing machine classification performance with new images based on the granular human biases captured.

Methods Our generator network was trained given uniform rather than Gaussian noise, which allows us to avoid proposing highly improbable stimuli to participants. Additionally, we avoid proposing states outside of this hypercube by forcing z to wrap around (proposals that travel outside of z are injected back in from the opposite direction by the amount originally exceeded). In particular, we run our MCMC chains through an unbounded state space by redefining each bounded dimension z_k as

$$z'_k = \begin{cases} -sgn(z_k) \times [1 - (z_k - \lfloor z_k \rfloor)], & \text{if } |z| > 1 \\ z_k, & \text{otherwise.} \end{cases} \quad (2)$$

Proposals were generated from an isometric Gaussian with a standard deviation of 0.1 60% of the time, and 0.7 otherwise.

We use this network to obtain large chains for two groups of five categories. Group 1 included *bottle*, *car*, *fire hydrant*, and

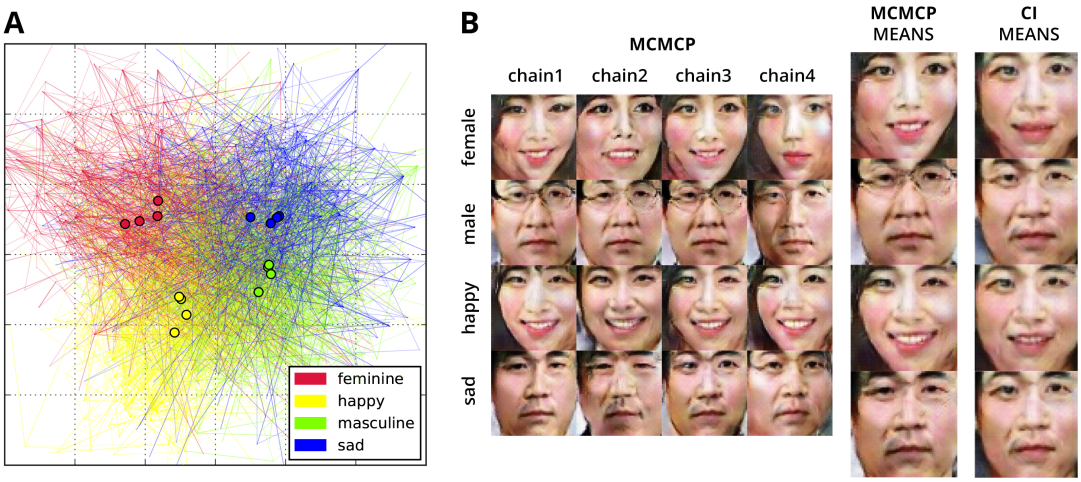

Figure 2: Visualizing captured representations. **A.** Fisher Linear Discriminant projections of all four MCMCP chains for each of the four face categories. The four sets of chains overlap to some degree, but are also well-separated overall. Means of individual chains are closer to other means from the same class than to those of other classes. **B.** Individual MCMCP chain means (4×4 grid) and overall category means (second to last) visualized as images (overall CI means also shown for comparison in the final column).

person, *television*, following Vondrick et al. (2015). Group 2 included *bird*, *body of water*, *fish*, *flower*, and *landscape*. Each chain was approximately $1,040$ states long, and four of these chains were used for each category (approximately $4,160$). In total, across both groups of categories, we obtained exactly $41,600$ samples from 650 participants.

To demonstrate the efficiency and flexibility of our method compared to alternatives, we obtained an equivalent number of trials for all categories using the variant of classification images introduced in Vondrick et al. (2015), with the exception that we used our BiGAN generator instead of the offline inversion previously used. This also serves as an important baseline against which to quantitatively evaluate our method because it estimates the simplest possible template.

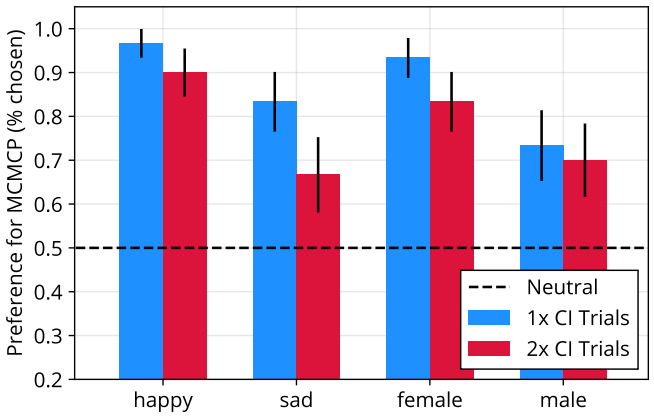

Figure 3: Human two-alternative forced-choice tasks reveal a strong preference for MCMCP means as representations of a category, when twice as many trials are used for CI.

Results The acceptance rate was approximately 50% for both category groups, which is near the common goal for MCMCP experiments. The samples for all ten categories are shown in Figure 5B and D using Fisher Linear Discriminant Analysis. Similar to the face chains, the four chains for each category converge to similar regions in space, largely away from other categories. In contrast, classification images shows little separation with so few trials (5C and D). Previous work suggests that at least an order of magnitude higher number of comparisons may be needed for satisfactory estimation of category means. Our method estimates well-separated category means in a manageable number of trials, allowing for the method to scale greatly. This makes sense given that CI compares arbitrary images, potentially wasting many trials, and clearly suffers from a great deal of noise.

Beyond yielding a decision rule, our method additionally produces a density estimate of the entire category distribution. In classification images, only mean template images can be viewed, while we are able to visualize several modes in the category distribution. Figure 4 visualizes these modes using the means of each component in a mixture of Gaussians density estimate. This produces realistic-looking multi-modal mental category templates, which to our knowledge has never been accomplished with respect to natural image categories.

Efficacy in classifying real images

Improvements of MCMCP over classification images may be both perceptible and detectable, but their practical differences are also worth considering — do they differ significantly on real-world tasks? Moreover, if the representations we learn through MCMCP are good approximations to people, we would expect them to perform reasonably well in categorizing real images. For this reason, we provide an additional quantitative assessment of the samples we obtained and

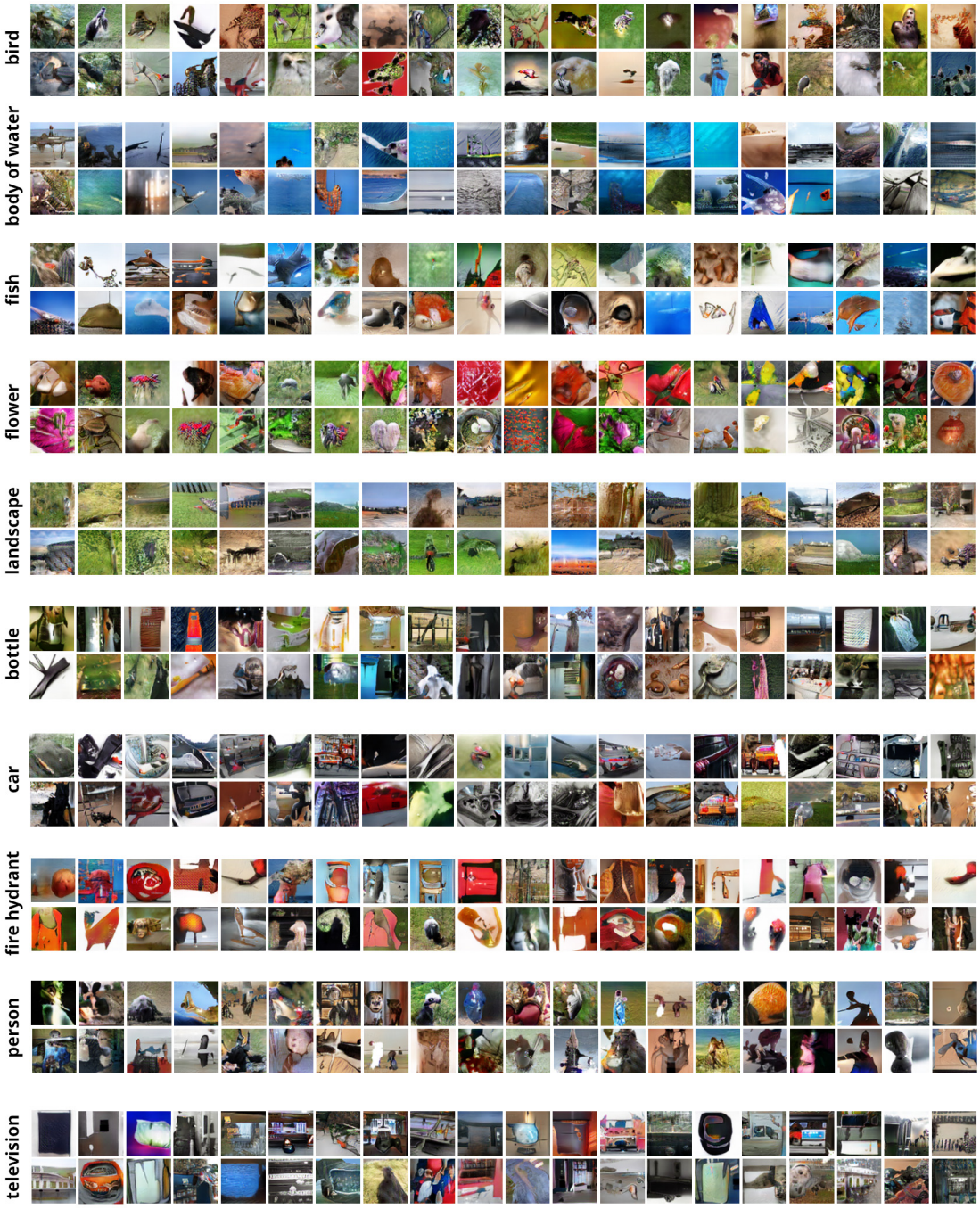

Figure 4: 40 most interpretable mixture component means (modes) taken from the 50 largest mixture weights for category.

compare them to classification images (CI) using an external classification task.

To do this, we scraped ≈ 500 images from Flickr for each of the ten categories, which was used for a classification task. To classify the images using our human-derived samples, we used (1) the nearest-mean decision rule, and (2) a decision rule based on the highest log-probability given by our ten density estimates. For classification images, only a nearest-mean decision rule can be tested. In all cases, decision rules based on our MCMCP-obtained samples overall outperform a nearest-mean decision rule using classification images (see Table 1). In category group 1, the MCMCP density performed best and was more even across classes. In category group 2, nearest-mean using our MCMCP samples did much better than a density estimate or CI-based nearest-mean.

Discussion

Our results demonstrate the potential of our method, which leverages both psychological methods and deep surrogate representations to make the problem of capturing human category representations tractable. The flexibility of our method in fitting arbitrary generative models allows us to visualize

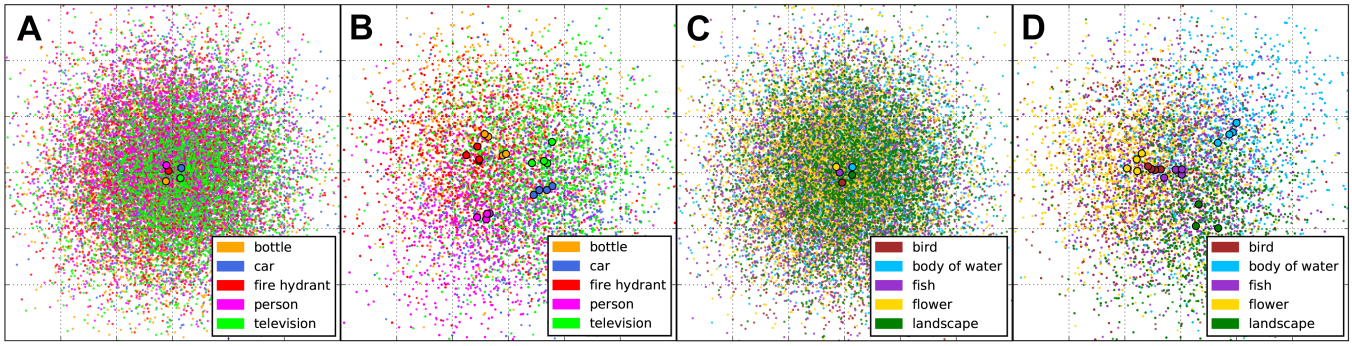

Figure 5: Categories are better separated by MCMCP representations. Fisher Linear Discriminant projections of **A**. CI comparisons for each category of group 1, **B.** samples for MCMCP chains for category group 1, **C.** CI comparisons for each category of group 2, and **D.** samples for MCMCP chains for category group 2. For A and C, large dots represent category means. For B and D, large dots represent chain means.

Table 1: Classification performance compared to chance for both category sets (chance is 0.20).

	bird	**body of water**	**fish**	**flower**	**landscape**	**all**
MM	.33	.28	.01	.57	.67	.37
MD	.23	.31	.18	.44	.73	**.38**
CM	.23	.30	.2	.24	.52	.30
	bottle	**fire hydrant**	**car**	**person**	**television**	**all**
MM	.15	.11	.32	.77	.73	**.42**
MD	.25	.26	.56	.19	.50	.35
CM	.28	.15	.62	.12	.13	.26

MM = MCMCP Mean, MD = MCMCP Density, CM = CI Mean

multi-modal category templates for the first time, and improve on human-based classification performance benchmarks. It is difficult to guarantee that our chains explored enough of the relevant space to actually capture the concepts in their entirety, but the diversity in the modes visualized and the improvement in class separation achieved are positive indications that we are on the right track. Further, the framework we present can be straightforwardly improved as generative image models advance, and a number of known methods for improving the speed, reach, and accuracy of MCMC algorithms can be applied to MCMCP make better use of costly human trials.

There are several obvious limitations of our method. First, the structure of the underlying feature spaces used may either lack the expressiveness (some features may be missing) or the constraints (too many irrelevant features or possible images wastes too many trials) needed to map all characteristics of human mental categories in a practical number of trials. Even well-behaved spaces are very large and require many trials to reach convergence. Addressing this will require continuing exploration of a variety of generative image models. We see our work as part of an iterative refinement process that can yield more granular human observations and inform new deep network objectives and architectures, both of which may yet converge on a proper, yet tractable model of real-world human categorization.

References

Ahumada. (1996). Perceptual classification images from vernier acuity masked by noise. *Perception*, *25*, 2–2.

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

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial networks. In *Advances in Neural Information Processing Systems* (pp. 2672–2680).

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representation (ICLR)*.

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).

Lake, B. M., 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.

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

Martin, J. B., Griffiths, T. L., & Sanborn, A. N. (2012). Testing the efficiency of Markov Chain Monte Carlo with people using facial affect categories. *Cognitive Science*, *36*(1), 150–162.

Nosofsky, R. M., Sanders, C. A., Meagher, B. J., & Douglas, B. J. (2017). Toward the development of a feature-space representation for a complex natural category domain. *Behavior Research Methods*, 1–27.

Peterson, J., Abbott, J., & Griffiths, T. (2016). Adapting deep network features to capture psychological representations. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*.

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Sanborn, A., & Griffiths, T. L. (2007). Markov Chain Monte Carlo with people. In *Advances in Neural Information Processing Systems* (pp. 1265–1272).

Vondrick, C., Pirsiavash, H., Oliva, A., & Torralba, A. (2015). Learning visual biases from human imagination. In *Advances in Neural Information Processing Systems* (pp. 289–297).