Deep Neural Networks Predict Category Typicality Ratings for Images

Brenden M. Lake
Center for Data Science
New York University

Wojciech Zaremba
Dept. of Computer Science
New York University

Rob Fergus
Dept. of Computer Science
New York University

Todd M. Gureckis
Dept. of Psychology
New York University

Abstract

The latest generation of neural networks has made major performance advances in object categorization from raw images. In particular, deep convolutional neural networks currently outperform alternative approaches on standard benchmarks by wide margins and achieve human-like accuracy on some tasks. These engineering successes present an opportunity to explore long-standing questions about the nature of human concepts by putting psychological theories to test at an unprecedented scale. This paper evaluates deep convolutional networks trained for classification on their ability to predict category typicality – a variable of paramount importance in the psychology of concepts – from the raw pixels of naturalistic images of objects. We find that these models have substantial predictive power, unlike simpler features computed from the same massive dataset, showing how typicality might emerge as a byproduct of a complex model trained to maximize classification performance.

Keywords: deep learning; neural networks; typicality; categorization; object recognition

Introduction

Recently, machine learning has made remarkable strides in developing systems that categorize objects. For most naturalistic images, especially those featuring a single object from a known class, the best algorithms can either correctly identify the object category or produce a series of plausible guesses. As part of the "deep learning" paradigm in machine learning, the largest recent advance in object categorization came from the AlexNet architecture (Krizhevsky, Sutskever, & Hinton, 2012), a massive convolutional neural network (convnet; LeCun et al., 1989) trained on 1.2 million raw pixel images to discriminate between 1000 different object categories. AlexNet won the 2012 ImageNet ILSVRC competition – the most challenging object categorization benchmark to date – by making approximately 40% fewer errors than the next best competitor. In the 2013 and 2014 ImageNet competitions, virtually all of the competitors used deep convnets at least partially inspired by the AlexNet architecture, furthering its advantages over alternatives such as hand-crafted computer vision features and other types of neural networks such as autoencoders and deep belief networks. Although it is difficult to directly compare human and machine performance on 1000-way classification, one estimate placed the best 2014 convnet (Szegedy, Liu, et al., 2014) only slightly behind human-level performance (Russakovsky et al., 2014).

These advances should interest the cognitive science community, especially since categorization is a foundational problem and some leading models are neural networks (e.g., Kruschke, 1992; Love, Medin, & Gureckis, 2004). Yet there has been little work evaluating the newest generation of neural networks as potential cognitive models or as large-scale tests of existing psychological theories. This paper offers a first step towards this goal by using convnets to predict human typicality ratings from raw naturalistic images.

Typicality ratings reflect the graded structure of concepts: people rate a Golden Retriever as a more typical "dog" than a hairless Chihuahua and a goldfish as a more typical "fish" than a shark. Since the seminal work of Rosch and colleagues (e.g., Rosch & Mervis, 1975), typicality has been a variable of paramount importance in the psychology of concepts. As Murphy (2002) puts it, for any task that requires relating an item to its category, typicality will influence performance, whether it is the speed of categorization, ease of production, ease of learning, usefulness for inductive inference, or word order in language. Previous work has found that typicality ratings can be predicted by human-produced feature descriptions (Rosch & Mervis, 1975) or similarity matrices (Ameel & Storms, 2006), but there have been no successful attempts in making predictions from raw naturalistic images.

However, there are reasons to suspect that convnets may not see the same typicality structure in images that people do, despite approaching human-level classification performance and predicting some aspects of neural response in monkey Inferior Temporal (IT) cortex (Yamins et al., 2014). First, the model parameters are trained strictly to optimize its ability to predict category labels, as opposed to predicting missing features or building a generative model of the data. It may be hard to learn prototypes with this objective: laboratory studies with human learners show that it discourages people from abstracting category prototypes when compared to feature prediction tasks (Yamauchi & Markman, 1998; Yamauchi, Love, & Markman, 2002). Second, recent work has shown it is easy to construct adversarial images that fool convnets but are easily recognized by people (Szegedy, Zaremba, et al., 2014). By examining the convnet's internal structure and modifying the image slightly, the model can be induced to mistake any image for any other category with an arbitrary degree of confidence. Nonetheless, these types of deformations must be rare occurrences in real images since the classifier generalizes well to unseen images.

If convnets predict human typicality, there would be implications for current psychological theories. In particular, it provides the opportunity to test existing theories using much harder problems at a much larger scale than typical laboratory studies (Griffiths, 2014), closer to the actual problems people face in the world. As mentioned in the paragraph above, training participants to predict missing labels rather than missing features discourages prototype formation in 2-way classification tasks. But 1000-way classification may not follow the same principles: it may be easier to learn 1000 prototypes (one for each class) rather than 499,500 discriminative rules

(one for each pair of classes), and thus large-scale simulations may offer new insights. Convnets also provide an opportunity to test for "contrast effects," the finding that objects are less typical if they resemble another category (Rosch & Mervis, 1975; Ameel & Storms, 2006), at a large scale by comparing different ways of extracting typicality from the model. Finally, by testing different models on the same massive dataset, we are able to explore classic questions of whether aspects of conceptual structure are bottom-up reflections of the world versus top-down impositions by the mind.

Methods

We asked people to rate a collection of images for category typicality, and we tested three convnet architectures and a baseline system on their ability to predict these ratings.

Stimuli. Typicality ratings were collected for eight categories from the ImageNet challenge: banana, bathtub, coffee mug, envelope, pillow, soap dispenser, table lamp, and teapot. They were chosen since they have high familiarity and a rich variation in typicality, unlike many of ImageNet's very specific categories such as "wire-haired fox terrier" or "ping-pong ball." We selected a set of 16 new images from each class that do not appear in the ImageNet training set (see Figs. 1 and 2 for examples), out of concern that photographs in the training set might be scored as typical because they are familiar to the network. Images were chosen via Google searches to span a maximum range of variation while focusing on a single, large, unoccluded object from a standard view.

Behavioral experiment. Human typicality ratings were collected on Mechanical Turk using 30 participants in the USA. Each participant rated every image from all 16 categories. After reading instructions from Rosch and Mervis (1975) Experiment 3, participants were asked "How well does this picture fit your idea or image of the category?" They responded from "1" (very good) to "7" (very poor). All 16 members of a category were presented sequentially, and participants viewed a grid of all of these images before beginning each category. They were paid $1.75, and the task (minus the instructions) took an average of 9.25 minutes (min = 4.5 and max = 20.5). A quiz checked for instruction comprehension and restarted the instructions if a question was missed.

Convolutional networks. We tested three different convnet architectures: OverFeat (Sermanet et al., 2014a), AlexNet (Krizhevsky et al., 2012), and GoogLeNet (Szegedy, Liu, et al., 2014). Pre-trained models were provided by the OverFeat ("fast model"; Sermanet et al., 2014b) and Caffe packages ("Reference CaffeNet" and "GoogLeNet"; Jia et al., 2014). While both OverFeat and GoogLeNet are derivatives of AlexNet, GoogLeNet is deeper and uses more sophisticated multi-resolution modules. We focus on OverFeat since it is particularly straightforward to describe.

OverFeat is a deep neural network with seven hidden layers. The first five hidden layers are convolutional, the last two hidden layers are standard fully-connected neural-network-style units, and the last layer is a 1000-way softmax layer, resulting in 145 million learned parameters and 2.8 billion connections. Convolutional layers take a set of 2D image-like grids as input (called "feature maps"), apply a set of trainable image filters, and output a new set of feature maps. The first two and the last convolutional layers also contain max pooling operations that reduce the resolution of the feature maps. Specifically, the model takes a 231x231 color image as input (three feature maps for RGB channels) and outputs 96 feature maps after applying 11x11 trainable image filters.[1] After three other layers of processing, the last convolutional layer has 1024 feature maps with smaller trainable filters (size 3x3). After the convolutions, the next two layers have 3072 and 4096 fully-connected connectionist units, respectively. Finally, the 1000-way softmax layer produces a probability distribution over the $j = 1, \ldots, 1000$ classes. It does so by first computing the raw class scores y_j from the activity x in the previous layer and weights w_{ij} and then computing the normalized class probabilities z_j, where

$$y_j = \sum_{i=1}^{4096} w_{ij} x_i \qquad \text{and} \qquad z_j = \frac{e^{y_j}}{\sum_{j=1}^{1000} e^{y_j}}. \tag{1}$$

The training objective is to maximize the log-probability of the correct label across the 1.2 million training instances (i.e., cross-entropy loss). The ImageNet dataset images were collected from search engines and verified on Mechanical Turk (Russakovsky et al., 2014).

An ensemble of multiple OverFeat models was entered in the ImageNet 2013 contest, where each trained model was identical but was initialize at a different random seed. The ensemble produced an top-five error rate of 14.2%, meaning that for over 85 percent of test images, the correct label appeared in the top five guesses. An ensemble AlexNet achieved an error rate of 16.4% in the 2012 contest and an ensemble GoogLeNet achieved 6.7% in 2014.

We assume that typicality ratings are related to the strength of a model's classification response to the category of interest. Ratings were extracted in two ways: either as the raw category score y_j ("raw typicality"; Eq. 1) or the normalized classification score $100 z_j$ ("contrast typicality"; Eq. 1) of the category of interest j (which may not be the model's largest response when considering all categories). Assuming the vector $w_{\cdot j}$ (Eq. 1) stores a prototype for category j, the raw score computes a measure of similarity (dot product) between the prototype and top-level hidden unit activations. In contrast, the normalized score more directly implements "contrast effects" as described in the Introduction, computing the raw score and then penalizing examples that score highly for other categories. Unfortunately, this is not an ideal test of

[1] Techniques exist for applying the model to rectangular images by averaging/maximizing across multiple square windows at different locations and scales. We side-stepped these complications by using square images cropped and rescaled to a model's desired input size with the main object approximately centered. As is standard for evaluating classification, typicality ratings were computed for an image and its mirror reflection, taking whichever value was higher.

contrast effects, since even the raw scores may show contrast effects due to the discriminative nature of the training. Both measures were evaluated.

Baseline SIFT model. We also tested a non-convnet baseline using code from the ImageNet 2010 challenge (Russakovsky et al., 2014). It is a standard computer vision pipeline of dense SIFT features (Lowe, 2004) quantized as a bag of 1000 visual words. Eight one-versus-all linear SVMs were trained – one for each category in the rating task – using all 1300 positive examples of these 8 classes and 100 randomly selected negative examples from each of the remaining 992 classes. SVM confidence was used to predict typicality.

Results and discussion

The mean typicality rating for each image was computed by averaging across participants. Spearman's rank correlation (ρ) was used to assess fit since human ratings were not expected to scale linearly with model ratings. First, the reliability of the human typicality ratings was assessed with a split-half correlation, which also serves as an approximate upper bound for model predictions. Across 25 random splits, the average reliability across all eight categories was $\rho = 0.92$, with "table lamp" as the most reliable ($\rho = 0.97$) and "soap dispenser" as the least ($\rho = 0.85$).

The convnets predicted human ratings about equally well regardless of whether raw or contrast typicality was used.[2] The full set of results for contrast typicality ratings is shown in Table 1. Across the eight categories, the mean rank correlation was $\rho = 0.67$ for OverFeat, $\rho = 0.67$ for AlexNet, $\rho = 0.63$ for GoogLeNet, and $\rho = 0.28$ for the SIFT baseline. A combination model that averages the predictions of the three convnets showed a slightly higher correlation of $\rho = 0.71$. It is worth noting that while we did not expect a linear relationship, the pearson correlations (r) were slightly higher (average Overfeat $r = 0.69$, AlexNet $r = 0.71$, GoogLeNet $r = 0.63$, Combination $r = 0.74$, and SIFT baseline $r = 0.27$). For the sake of completion, the average correlation for raw typicality ratings was $\rho = 0.65$ ($r = 0.68$) for OverFeat, $\rho = 0.67$ ($r = 0.69$) for AlexNet, $\rho = 0.69$ ($r = 0.72$) for GoogLeNet.

Typicality ratings from people and OverFeat are shown for five categories in Figs. 1 and 2, offering some insight into the differences. While illustrated for OverFeat, these differences are evident in the other convnets. For bananas (Fig 1), people may have ranked the images based on their similarity to an "ideal" (Barsalou, 1985); in this case, a yellow spot-free banana. In contrast, OverFeat rated a greenish plantain and a spotted banana about as highly as more ideal bananas, raising the possibility that this may be more a top-down imposition from the mind rather than a bottom-up property of visual experience with bananas (most bananas are not perfect). For envelopes, there appears to be similar ideal based on standard white envelopes that is reflected more strongly in the human ratings. For pillows, people rated rectangular bed pillows as more typical than decorative couch pillows, while OverFeat showed the opposite pattern, perhaps due to a curious paucity of bed pillows in the ImageNet training set. Finally, some of the outliers were images for which the model preferred a different class, including the red bathtub (mislabeled a dining table) and the blue coffee mug (a bucket/pail).

[2]The scale was reversed for the human ratings (1 to 7) so that larger values are more typical.

Table 1: Rank correlations for human and machine typicality.

Category	OverFeat	AlexNet	GoogLe	Combo	SIFT
Banana	0.82	0.8	0.73	0.84	0.4
Bathtub	0.68	0.74	0.48	0.78	0.39
Coffee mug	0.62	0.84	0.84	0.85	0.63
Envelope	0.79	0.62	0.75	0.78	0.38
Pillow	0.67	0.55	0.69	0.59	0.11
Soap Disp.	0.74	0.79	0.82	0.75	0.09
Table lamp	0.69	0.8	0.7	0.83	0.48
Teapot	0.38	0.21	0.07	0.28	-0.23
Average	**0.67**	**0.67**	**0.63**	**0.71**	**0.28**

Our results suggest that deep convnets learn graded categories that can predict human typicality ratings, at least for some types of everyday categories. Outside of this work, few studies have tried to predict high-level cognitive measures from the pixels of naturalistic images, making it difficult to compare the size of these correlations with past work. One study by Isola, Xiao, Parikh, Torralba, and Oliva (2013) showed that image memorability (which is not necessarily analogous to typicality) could be predicted from raw images with a rank correlation of $\rho = 0.46$ after training a model directly on memorability data. Not only are our correlations stronger, the models were not trained to predict typicality at all. Not surprisingly, the convnets have lower predictive power for typicality than models receiving processed input data such as human feature ratings or similarity ratings that usually produce correlations greater than $\rho \geq 0.80$ (e.g., Rosch & Mervis, 1975; Ameel & Storms, 2006).

A limitation of our results is that the relationship between classifier performance and typicality effects remains unclear, making it difficult to isolate any unique contributions of the architectures beyond their abilities as classifiers. The low correlations from the SIFT baseline suggest that human typicality ratings are not just a property of any classifier trained on a large dataset with reasonable features. It is also worth noting that GoogLeNet, although superior for object recognition, was not better at predicting human typicality. But we cannot yet compare against equally high-performance computer vision systems that operates by different principles, since these models do not yet exist.

The role of feature complexity. To gain further insight into how the convnets predict typicality, we analyzed the structure present at each layer of processing. Since features increase in complexity and category specificity with depth (e.g., Zeiler & Fergus, 2014; Yosinski, Clune, Bengio, & Lipson, 2014), the depth at which typicality emerges suggests the difficulty of extracting this structure from the raw data. To make predictions at intermediate layers, the hidden layer activations

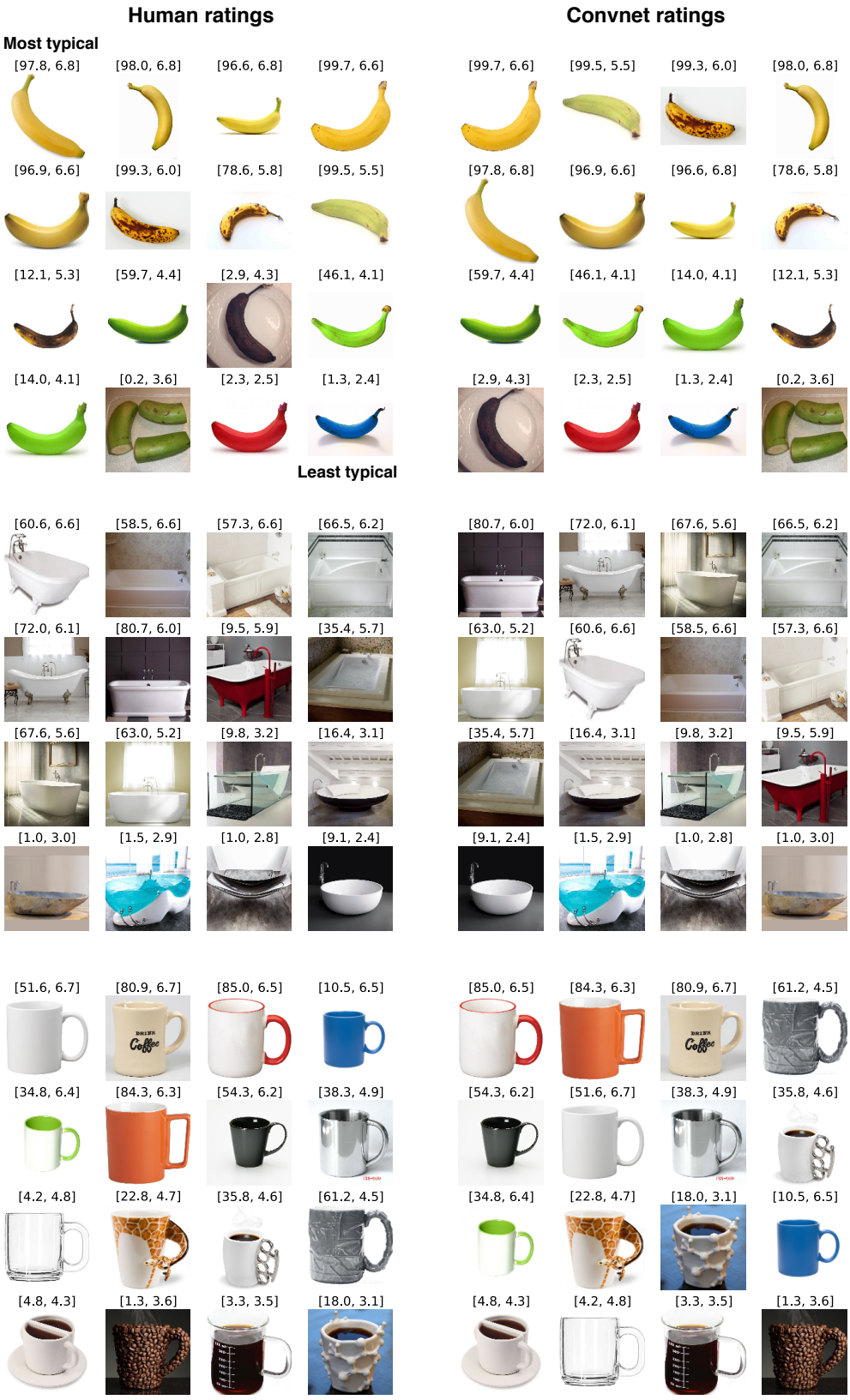

Figure 1: Images ranked from most to least typical by people (left) and the OverFeat convnet (right). Rankings flow left to right and then top to bottom. The values above each image $[x_1, x_2]$ show the convnet contrast typicality rating and the mean participant rating, respectively. The categories include banana, bathtub, and coffee mug.

Figure 2: Images ranked from most to least typical. See caption from Fig. 1. The categories include envelope and pillow.

(pre-pooling) were extracted for all 1300 training images of each category (center-cropped). For each layer, the average activation vector was computed for each class to serve as the category prototype. Typicality was modeled as the cosine distance between the activation vector for a new image and the stored prototype. For the top layer, the correlation with human ratings was the same as our previous results (average $\rho = 0.67$ for OverFeat and AlexNet; GoogLeNet was not analyzed). Performance steadily improves with depth (Fig. 3), again confirming that typicality does not automatically emerge from a large dataset with simple feature extraction. The data must be viewed through the right lens before the structure is apparent.

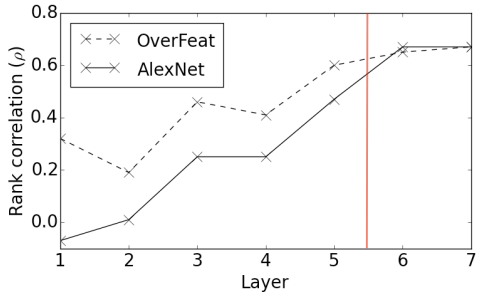

Figure 3: Correlation between human and convnet typicality ratings as a function of network depth. The red line indicates a transition from convolutional (1-5) to standard layers (6-7).

Conclusions

This paper evaluated the ability of convolutional neural networks (convnets) to predict human typicality ratings for images of category examples – a critical variable that influences performance in nearly all tasks that involve categorical processing (Murphy, 2002). These models were trained only to predict category labels, and despite previous human studies on 2-way categorization suggesting that this tasks promotes the extraction of discriminating features rather than prototypes (Yamauchi & Markman, 1998; Yamauchi et al., 2002), convnets trained on 1000-way classification were able to pre-

dict human typicality ratings with an average rank correlation of 0.67 (OverFeat and AlexNet) or 0.63 (GoogLeNet). Different operationalizations of typicality provided equally good fits, suggesting there was no particular benefit for an explicitly contrastive measure of typicality (Rosch & Mervis, 1975; Ameel & Storms, 2006). Additional analyses explored the role of the training data versus the model in capturing typicality, finding that simple features did not provide good prototypes for prediction even with many examples per class. Finally, convnets were less sensitive to category ideals than people, suggesting that feature extraction on a large dataset may not be fully sufficient for ideals to arise.

This is just a first step towards understanding the "synthetic psychology" of deep neural networks and mining them for insights about human conceptual structure. We tested only pretrained systems, leaving questions about learning and development for future research. Further studies could test whether convnets show faster learning of categories that are separable on one dimension (e.g., Shepard, Hovland, & Jenkins, 1961), faster learning of categories with mostly typical examples (Posner & Keele, 1968), or a preference for learning typical examples first (Rosch, Simpson, & Miller, 1976) – insights that could inspire new training procedures for deep learning. Additional studies could test for a coarse-to-fine pattern of category differentiation (Rogers & McClelland, 2004) or study the typicality of higher-level categories such as "dog" or "furniture." Finally, convnet activations have been shown to predict neural response in monkey IT cortex, where both systems show higher similarity within and lower similarity between categories (Yamins et al., 2014). Given our results, it may also be promising to use these methods to study more fine-grained structure within categories.

Whether or not convnets can match these aspects of behavior, they are still far too limited compared to the human ability to learn and use new concepts. While the convnet was trained on an average of 1200 images per class, people need far less data in order to learn a new category (Lake, 2014). In addition, human concepts support the flexible use of the same knowledge across many tasks – classification, inference, generation, and explanation – a remarkable quality that current machine learning approaches do not capture. While the current best algorithms are limited compared to people, further exercises in understanding their synthetic psychology may serve to both advance machine learning and psychological theory.

Acknowledgments. We thank the Moore-Sloan Data Science Environment at NYU for supporting this work.

References

Ameel, E., & Storms, G. (2006). From prototypes to caricatures: Geometrical models for concept typicality. *Journal of Memory and Language*, *55*(3), 402–421.

Barsalou, L. W. (1985). Ideals, Central Tendency, and Frequency of Instantiation as Determinants of Graded Structure in Categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(4), 629–649.

Griffiths, T. L. (2014). Manifesto for a new (computational) cognitive revolution. *Cognition*, 10–12.

Isola, P., Xiao, J., Parikh, D., Torralba, A., & Oliva, A. (2013). What makes a photograph memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(7), 1469–1482.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., ... Eecs, U. C. B. (2014). Caffe: Convolutional Architecture for Fast Feature Embedding. *ACM Conference on Multimedia*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44.

Lake, B. M. (2014). *Towards more human-like concept learning in machines: Compositionality, causality, and learning-to-learn*. Unpublished doctoral dissertation, MIT.

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*, 541–551.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A Network Model of Category Learning. *Psychological Review*, *111*(2), 309–332.

Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.

Murphy, G. L. (2002). *The Big Book of Concepts*. Cambridge, MA: MIT Press.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*(3), 353–363.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic Cognition*. Cambridge, MA: MIT Press.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605.

Rosch, E., Simpson, C., & Miller, R. S. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance*, *2*(4), 491–502.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2014). *ImageNet Large Scale Visual Recognition Challenge*.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & LeCun, Y. (2014a). OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. In *International Conference on Learning Representations (ICLR 2014)*.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & LeCun, Y. (2014b). *OverFeat: Object Recognizer, Feature Extractor*. Retrieved from `http://cilvr.nyu.edu`

Shepard, R. N., Hovland, C. L., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, *75*(13, Whole No. 517).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2014). *Going Deeper with Convolutions*.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR 2014)*.

Yamauchi, T., Love, B. C., & Markman, A. B. (2002). Learning nonlinearly separable categories by inference and classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 585–593.

Yamauchi, T., & Markman, A. B. (1998). Category Learning by Inference and Classification. *Journal of Memory and Language*, *39*(39), 124–148. doi: 10.1006/jmla.1998.2566

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. a., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–24.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems (NIPS)*.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. In *European Conference on Computer Vision (ECCV)*.