CogSci 2018 Paper

Reuben Feinman (reuben.feinman@nyu.edu)
Center for Neural Science
New York University

Brenden M. Lake (brenden@nyu.edu)
Department of Psychology and Center for Data Science
New York University

Abstract

People use rich prior knowledge about the world in order to efficiently learn new concepts. These priors–also known as "inductive biases"–pertain to the space of internal models considered by a learner, and they help the learner make inferences that go beyond the observed data. A recent study found that deep neural networks optimized for object recognition develop the shape bias (Ritter et al., 2017), an inductive bias possessed by children that plays an important role in early word learning. However, these networks use unrealistic training data, and the conditions required for these biases to develop are not well understood. Moreover, it is unclear whether the learning dynamics of these networks bear any relationship to developmental processes in children. We investigate the development and influence of the shape bias in neural networks using controlled datasets of abstract patterns and synthetic images, allowing us to systematically vary the quantity and form of the experience provided to the learning algorithms. We find that simple neural networks develop a shape bias after seeing as few as 3 examples of each concept, and that these biases tend to strengthen with depth in the network. The development of these biases predicts the onset of vocabulary acceleration in our networks, consistent with the developmental process in children.

Keywords: neural networks; inductive biases; learning-to-learn

Introduction

Humans possess the remarkable ability to learn a new concept from seeing just a few examples. A child who is learning her first few words can easily pick up the meaning of the word "fork" after observing only one or a handful of forks (Bloom, 2000). In contrast, state-of-the-art artificial learning systems use hundreds or thousands of examples when learning to recognize the same objects (e.g., Krizhevsky et al., 2012; Szegedy et al., 2015). Consequently, significant effort is ongoing to understand what neural and cognitive mechanisms enable efficient concept learning (Lake et al., 2017). In this paper, we perform a series of developmentally-informed neural network experiments to study the computational basis of efficient word learning. [1]

If humans extrapolate beyond the presented data, then another source of information must make up the difference; prior background knowledge must delimit the hypothesis space during learning (Tenenbaum et al., 2011; Lake et al., 2017). By constraining the space of models considered by the learner, these priors, referred to herein as "inductive biases," help the learner make inferences that go far beyond the observed data. As one manifestation, human children make use of the shape bias–the assumption that objects that have the same name will tend to have the same shape–when learning new object names, and thus shape is more important than color, material and other properties when generalizing a new label to new examples (Fig. 1) (Landau et al., 1988). Similarly, children assume that object names are mutually exclusive, i.e. that a novel name probably refers to a novel object rather than a familiar object (Markman & Wachtel, 1988). Although the origin of inductive biases is not always clear, results show that children, adults and primates can "learn-to-learn" or form higher-order generalizations that can improve the efficiency of future learning (Harlow, 1949; Smith et al., 2002; Dewar & Xu, 2010).

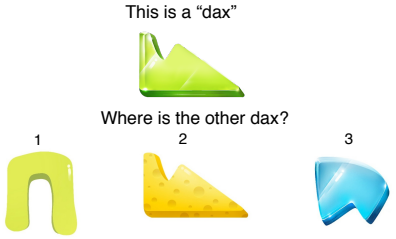


Figure 1: The shape bias. Children learn that objects with the same name tend to have the same shape, and thus option 2 above is likely the right answer. This inductive bias helps with future word learning.

Cognitive scientists have proposed a number of computational models to explain how inductive biases are acquired and harnessed for future learning. Hierarchical Bayesian Models (HBMs) enable probabilistic inference at multiple levels simultaneously, allowing the model to learn the structure of individual concepts while simultaneously learning about the structure of concepts in general (i.e., learning a prior on new concepts) (Gelman et al., 2013; Kemp et al., 2007; Salakhutdinov et al., 2012). These models have been used to explain various forms of "learning-to-learn" including learning a shape bias (Kemp et al., 2007), yet it is currently difficult to apply HBMs to the type of raw, high-dimensional sensory data that children receive, such as images or audio waves. In some cases, HBMs and related approaches have been applied successfully to raw high-dimensional data such as when learning handwritten characters (Lake et al., 2015), but with the help of domain-specific knowledge and engi-

[1] All experiments can be reproduced using the code repository located at `http://github.com/rfeinman/learning-to-learn`.

neering. In contrast, recent progress in training deep neural networks shows how relatively generic architectures can learn effectively from raw data (LeCun et al., 2015), providing the potential bridge between controlled simulations with synthetic data (e.g., Colunga & Smith, 2005) and large-scale real-world object recognition tasks with raw data (e.g., Ritter et al., 2017). Here, we take advantage of this connection by using neural networks to study learning-to-learn in several different settings of varying stimulus complexity, with the goal of isolating the fundamentals of the learning dynamics.

Most related to our work here are studies by Colunga & Smith (2005) and Ritter et al. (2017) investigating neural network accounts of shape bias development. Colunga & Smith (2005) showed that a simple neural network model, trained via Hebbian learning, can acquire the shape bias when presented with datasets comparable to human developmental studies. However, these simulations operate on highly simplified bit-vector data, and it is unclear how their results generalize to more realistic stimuli. Furthermore, the authors did not systematically vary the structure of the training set, so we do not know the exact conditions in which biases arise, nor whether current models are sufficient to explain it. In a recent study, building on recent advances in object recognition, Ritter et al. (2017) found that performance-optimized deep neural networks (DNNs) develop the shape bias over the course of learning when trained on the popular ImageNet classification dataset consisting of raw naturalistic images. Although these results highlight an exciting possible connection between DNNs and developmental psychology, many questions remain. ImageNet–which contains thousands of labeled examples of each visual concept–is a poor proxy for human developmental learning sets. Whether or not these models can acquire the same bias with a training set comparable to humans remains unknown; an answer to this question may help to explain the developmental process in children. Furthermore, while the development of the shape bias is known to predict the onset of vocabulary acceleration in children (Gershkoff-Stowe & Smith, 2004), we do not know whether the same holds for DNNs.

We investigate the development and influence of inductive biases in neural network models using artificial object stimuli designed to closely mimic developmental studies with human children. Borrowing the experimental paradigm of Smith et al. (2002), we evaluate the first- and second-order generalization capabilities of neural networks trained with variable-sized datasets. Beginning with simple bit-vector data akin to Colunga & Smith (2005), we systematically vary the number of categories and the number of exemplars in the training set, recording generalization performance at each pairing. Parallel experiments are then performed with RGB image data, where each image consists of a 2D object with a particular shape, color and texture that is shifted and placed over white background. For each the bit-vector and RGB image data, we investigate the parametric relationship between bias strength and attribute similarity in our models by systematically varying the shape, color and texture attributes of select test stimuli. Similarly, we evaluate bias strength as a function of depth in the network. In a final set of experiments, we investigate the correlation between shape bias acquisition and the rate of concept learning in our networks, mirroring an analogous study from human developmental psychology (Gershkoff-Stowe & Smith, 2004).

Experimental Paradigm

We first set out to model the infant learning tasks described by Smith et al. (2002) using simple neural networks. During the study, 17-month-old children were taught the names of primitive objects over the course of 7 weeks via weekly play sessions. Objects in the study were 3D formations constructed of various materials; each object contained a specific shape, color and texture (material), and the names of the objects were organized stricly by shape. During the weekly sessions, children played around with each object while an adult announced the name of the object that they were playing with repeatedly. By the end of the study, the children subjects had learned the shape bias–i.e., they had formed the generalization that only objects with the same shape have the same name. A control group of children, whom did not partake in the play sessions, did not form the same generalization.

To model this study, we use artificial object datasets designed to mimic the training data presented to the children subjects. We first perform our computational experiments with with categorical bit-vector data, followed by RGB images (the succeeding two sections, respectively). Each object sample is assigned a shape, texture and color value. We train simple neural network models to classify the shape of each object, providing labels that mimick those provided to the children. To evaluate the generalization capabilities of our models post-training, we use two generalization tests modeled after the two tests used by Smith et al. (2002) to assess the children after training.

1. First-order generalization test: For the first-order generalization test, infants are first presented with a baseline object that they have seen and played with during training. Then, they are presented with three novel objects that have not been seen before: one that matches the baseline in shape, one that matches in color, and one that matches in texture. For each of the three, the other two feature dimensions are novel. The infants are asked to select which of the three comparison objects share the same name as the baseline. Performance is measured as the fraction of trials in which the child selected the correct object, i.e. the shape match. We simulate this test by creating an evaluation set containing groupings of four samples: a baseline, a shape match, a color match, and a texture match. We find which of the three samples the network thinks to be most similar by evaluating the cosine similarity [2] using features at the highest hidden layer of the model. Accuracy is defined as the fraction of groupings for which the

[2]Near-identical results were observed using Euclidean distance.

model chose the correct (shape-similar) object. This test was repeated for different training set sizes, i.e. different combinations of $\{\#\ categories, \#\ exemplars\}$.

2. Second-order generalization test: For the second-order generalization test, infants are first presented with a baseline object that is novel in shape, color and texture. From there, the trial proceeds similarly to those of the first-order test: a shape match, color match and texture match are presented, and the child must select which object she believes to share a name with the baseline. All shapes, colors and textures are novel to the child in this test. We simulate this test by creating an evaluation set... TODO. This test was repeated for different training set sizes, i.e. different combinations of $\{\#\ categories, \#\ exemplars\}$.

Multilayer Perceptron Trained on Synthetic Objects

To begin with, we use a simple multilayer perceptron (MLP) that operates on categorical data. Since shape, color, & texture have categorical feature values, we encode the values using unique bit vectors that are randomly assigned at the beginning of the experiment. We use a simple feed-forward NN with one hidden layer of 30 units, and the ReLU activation function. The number of units in the softmax layer depends on the *# categories* parameter for the particular dataset. Results are shown in Fig. **??**.

Parametric Tests

Blah.

Convolutional Network Trained on Synthetic Objects

As a second type of model, we used a simple convolutional neural network (CNN) architecture consisting of... TODO. To train this CNN, we generated images of artificial 2-D objects (Fig. 4) with a variety of different shape, color and texture values. TODO... finish. Results are shown in Fig. **??**.

Layer-wise Biases

The first step of our analysis is to evaluate the shape, color and texture biases of VGG-16 at each of its layers, in order to get a picture of how these biases develops along the higherarchy of the model's internal representation. In order to probe the model, we make use of two unique image datasets with stimuli that mimic Smith et al. (2002).

1. Artist-generated object dataset: These images were generated by an artist in Adobe Photoshop. See Fig. 5.

2. CogPsyc object dataset: These images were provided by cognitive psychologist Linda Smith, and they were used in the experiments of Ritter et al. (2017). See Fig. 6.

The layer-wise bias results are shown in Fig. 7a.

Parametric Tests

The second step of our analysis is to examine how the shape and color biases depend on the intensity of their respective feature similarities. For these tests, we make use of our computer-generated images so that we can quantitatively manipulate and evaluate the shape and color features of our objects. The parametric bias results are shown in Fig. 8 and 9.

Predicting the Onset of Vocabulary Acceleration

Acknowledgements

This research was supported by a Henry M. McCracken fellowship and NYU start-up faculty funding? We thank Subhankar Ghosh for useful code and preliminary experiment ideas.

References

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

Colunga, E., & Smith, L. B. (2005). From the lexicon to expectations about kinds: a role for associative learning. *Psychological Review*, *112*(2), 347–382.

Dewar, K. M., & Xu, F. (2010). Induction, overhypothesis, and the origin of abstract knowledge: evidence from 9-month-old infants. *Psychological Science*, *21*(12), 1871-1877.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2013). *Bayesian data analysis*. Boca Raton, FL: Chapman & Hall.

Gershkoff-Stowe, L., & Smith, L. B. (2004). Shape and the first hundred nouns. *Child Development*, *75*(4), 1098–1114.

Harlow, H. F. (1949). The formation of learning sets. *Psychological Review*, *56*(1), 51–65.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical bayesian models. *Developmental Science*, *10*(3), 307–321.

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems 25* (pp. 1097–1105).

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, *3*(3), 299–321.

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

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meaning of words. *Cognitive Psychology*, *20*(2), 121–157.

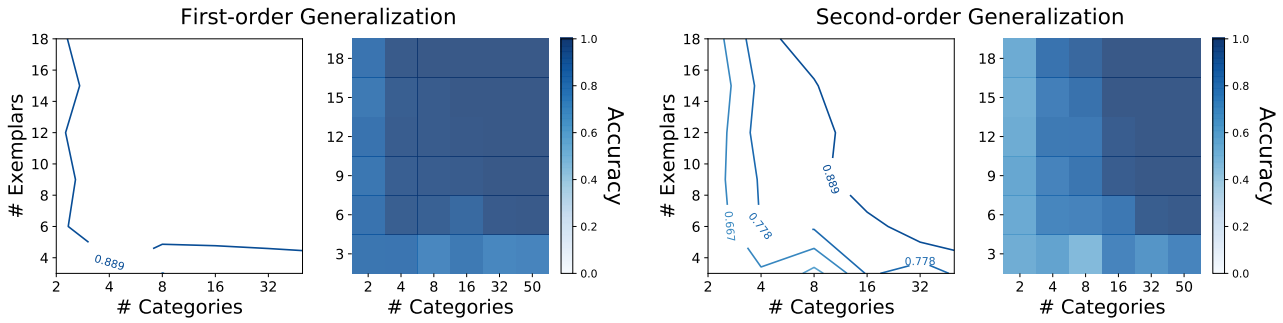


Figure 2: MLP generalization results for various training set sizes. Results show the average from 10 trials.

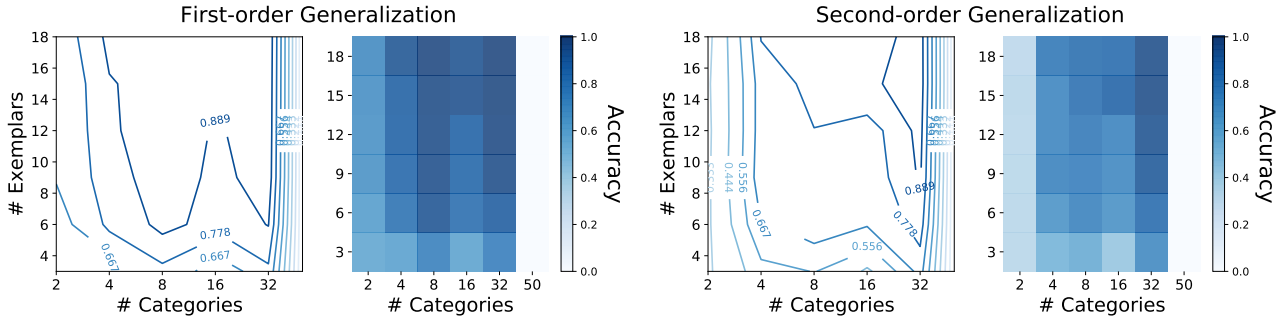


Figure 3: CNN generalization results for various training set sizes. Results show the average from 10 trials.

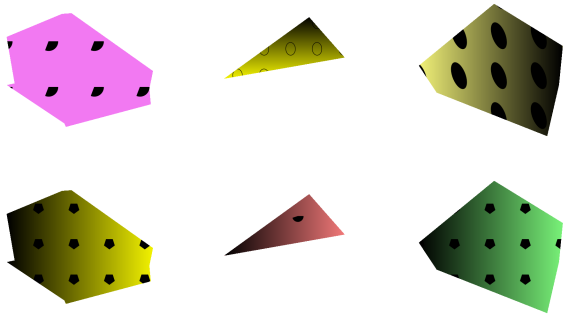


Figure 4: Computer-generated images of 2D objects with different shape, color and texture features.

Figure 5: Artist-designed images of 3D objects with different shape, color and texture features.

Figure 6: CogPsyc images with different shape, color and texture features (TODO).

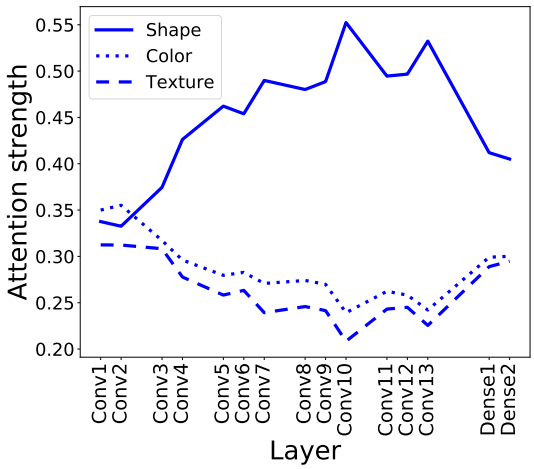
Ritter, S., Barrett, D. G. T., Santoro, A., & Botvinick, M. M. (2017). Cognitive psychology for deep neural networks: a shape bias case study. In *Proceedings of the 34th international conference on machine learning* (pp. 2940–2949).

Salakhutdinov, R., Tenenbaum, J., & Torralba, A. (2012). One-shot learning with a hierarchical nonparametric bayesian model. In *Proceedings of the icml workshop on unsupervised and transfer learning* (Vol. 27, pp. 195–206).

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, *13*(1), 13–19.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the 2015 ieee conference on computer vision and pattern recognition* (p. 1-9).

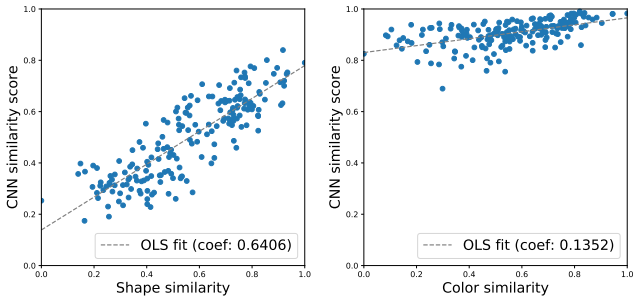
Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.



(a) Artist-generated images

(b) CogPsyc images (TODO)

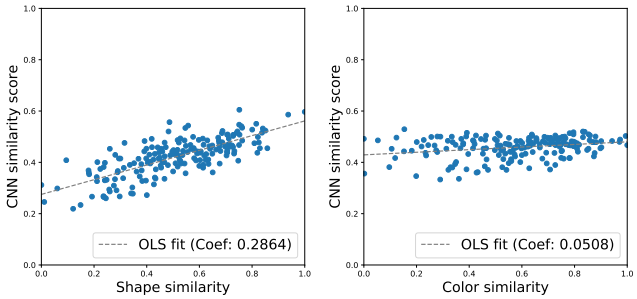
Figure 7: VGG-16 layer-wise biases on two image datasets. Attention strength refers to the network's similarity score between the target objectand objects that match in either shape, color or texture.



(a) Shape

(b) Color

Figure 8: VGG-16 parametric shape and color biases w/ other features constant.



(a) Shape

(b) Color

Figure 9: VGG-16 parametric shape and color biases w/ other features varying.