Learning Inductive Biases with Developmentally-informed Neural Networks

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 they attend to shape more often 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).

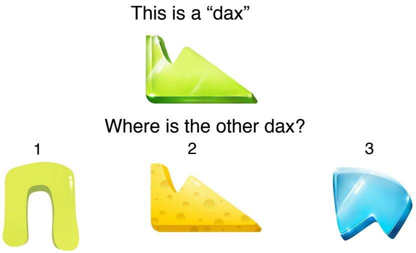


Figure 1: The shape bias. Children learn that objects with the same name tend to have the same shape, and thus option 2 above is likely the right answer. This inductive bias helps with future word learning.

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. Additionally, 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 mimick the training data presented to the children subjects. We first perform our computational experiments with categorical bit-vector data, followed by RGB images. The details of these two data formats and their corresponding network architectures are described in the succeeding two sections, respectively. Each object sample is assigned a shape, texture and color value. We train simple neural networks to classify the shape of each object, providing labels that mimick those provided to the children. Training is performed for various dataset sizes, varying both the number of categories and the number of exemplars of each category provided to the network. We evaluate the generalization capabilities of the network for each training set size using two generalization tests modeled after the two tests of Smith et al. (2002).

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 match, the other two stimulus aspects 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. Smith et al. (2002) propose that children who display the first-order generalization capability have taken the first step in the development of the shape bias: they have learned to attend to shape when

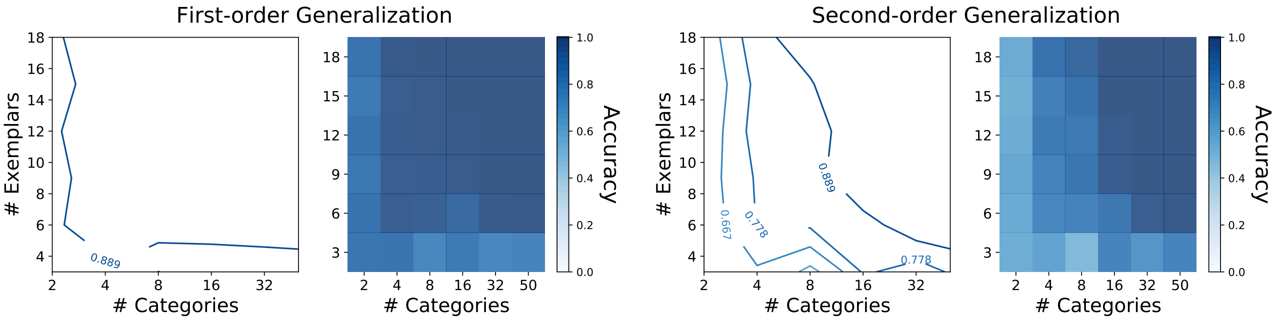


Figure 2: MLP generalization results for various training set sizes. Results show the average of 10 training runs.

identifying exemplars of familiar object categories. To simulate this test, we create 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 last hidden layer of the model. Accuracy is defined as the fraction of groupings for which the model chose the correct (shape-similar) object.

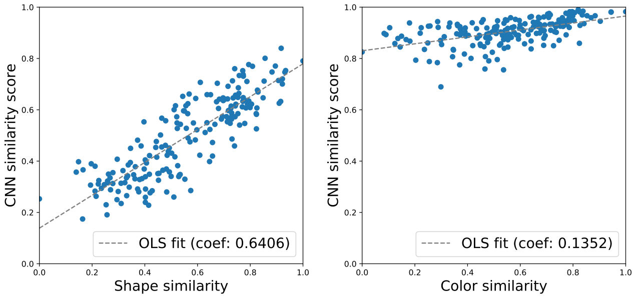
2. Second-order generalization test: For the second-order generalization test, infants are first presented with a baseline object that is novel in shape, color and texture. From there, the trial proceeds similarly to those of the first-order: a shape match, color match and texture match are presented, 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. Smith et al. (2002) propose that children who display the second-order generalization capability have taken one step further beyond the first-order test in the development of the shape bias. These children have not only learned to categorize a handful of object cateogories by shape, but they have induced that shape is a useful feature in general when categorizing objects. This induction, it was shown, is useful for vocabulary development. We simulate the second-order test with artificial object stimuli similarly as done in the first-order case, again using last hidden layer features to quantify model similarity scores.

We hold a novel set of shapes, colors and textures to be used for the generalization tests in our experiments. Accuracy over 2000 test trials is recorded as the performance metric for each generalization.

Multilayer Perceptron Trained on Synthetic Objects

In our first experiment, we encode the categorical stimulus information of our synthetic objects as abstract patterns, representing the shape, color and texture features each as independent 20-bit vectors. The overall stimulus thus has 60 bits. The categories of each feature are assigned their bit-vector representations randomly at the beginning of the experiment. We train a multilayer perceptron (MLP) with one hidden layer of 30 ReLU units to classify the shape of the presented stimulus. The number of units in the softmax layer varies with the *# categories* dataset parameter. To mimick the control group of Smith et al. (2002), we initialized our MLP model randomly and evaluated second-order generalization results prior to training. The model selected by shape X%, color X% and texture X% We then trained our model with various dataset sizes. Results for each the first- and second-order generalizations, averaged over 10 training runs, are shown in Fig. 2. TODO: discuss results.

To analyze the parametric dependency of network biases on the presented stimuli, we perform a series of tests using a network trained with 50 categories and 15 exemplars. For the first test, we probe the shape bias of our MLP by varying the shape distance between two presented stimuli and recording the resulting network similarity score at each input pair, using cosine similarity at the hidden layer. Distance in shape space is quantified as the fraction of bits that differ along the 20-bit space denoted to shape. Similar tests are also performed for the color and texture features. Results are shown in Fig. 3. TODO: discuss results.



(a) Shape

(b) Color

Figure 3: MLP parametric shape and color biases. TODO: put correct result plots here.

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

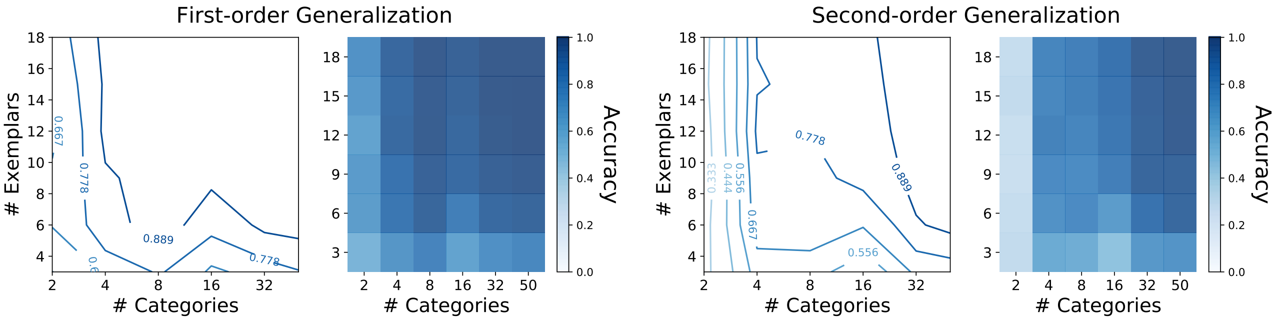


Figure 4: CNN generalization results for various training set sizes. Results show the average of 10 training runs.

Convolutional Network Trained on Synthetic Objects

Our first experiment uses a highly simplified data format. However, neural network architectures can perform well with complex data, including natural images, language and audio. In a successive experiment, ask whether similar generalization results can be achieved using object stimuli encoded as RGB images. To do so, we construct an image dataset consisting of artifical 2D objects. Each object is a 2D shape of a specified color placed over white background. Objects are initially centered, however, a random shift is applied to every sample. Texture is represented in a fourth dimension, independent of RGB space. The motivation for this design choice is as follows. In the experiments of Smith et al. (2002), children physically touch each object that they are presented in addition to observing it visually. Although the presence of materials like plastic and styrofoam may be visually subtle compared to color and shape, these materials become much more detectable when sensed by hand. Since visual and touch signals are received along separate pathways, it makes sense to provide these signals along independent axes to a computational model.

Object shapes are determined by randomly generating sets of points around the 2D image window, and colors are generated to span the RGB vector space with even separation. We use black & white textures from the Brodatz database (Brodatz, 1966) for our texture categories. Some examples of our objects are shown in Fig. 5. We train a convolutional neural network (CNN) consisting of two convolution layers with five filters, each followed by a max pooling layer. The last pooling layer is followed by a fully-connected layer of 25 ReLU units, and the softmax layer again varies in size according to the number of categories in the dataset. The initial second-order generalization results for a random network are as follows: shape 0.39, color 0.41 and texture 0.20. Results for networks trained on various dataset sizes are shown in Fig. 4.

We next analyze the parametric relationship between our CNN biases and the feature values of presented stimuli similarly to our MLP. For our image objects, distance in shape space is quantified as the Modified Hausdorff Distance (Dubuisson & Jain, 1994) between the shape pair. In color space, distance is quantified using the cosine similarity of the RGB vector pair. The parametric bias results are shown in Fig. 6.

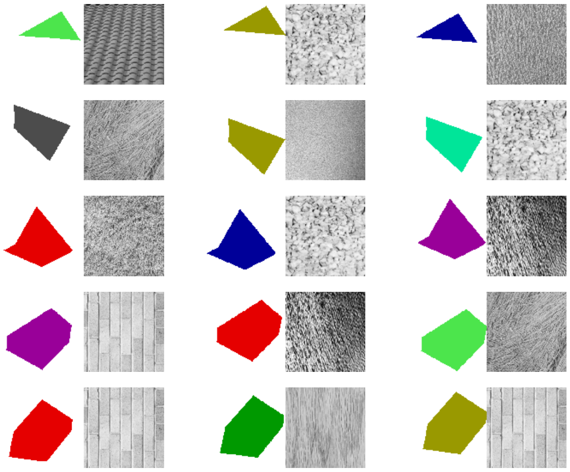
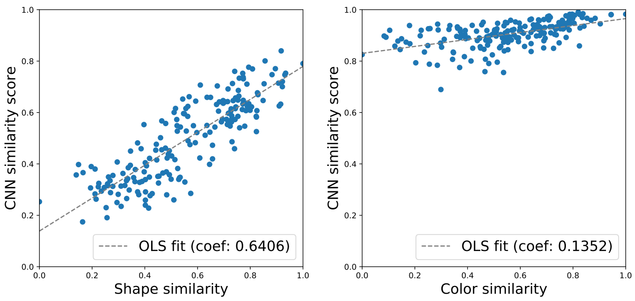


Figure 5: Computer-generated images of 2D objects with different shape, color and texture features. Rows correspond to object categories. Dimensions 1-3 are shown on the left of each image, and dimension 4 (texture) is shown on the right.



(a) Shape

(b) Color

Figure 6: CNN parametric shape and color biases. TODO: put correct result plots here.

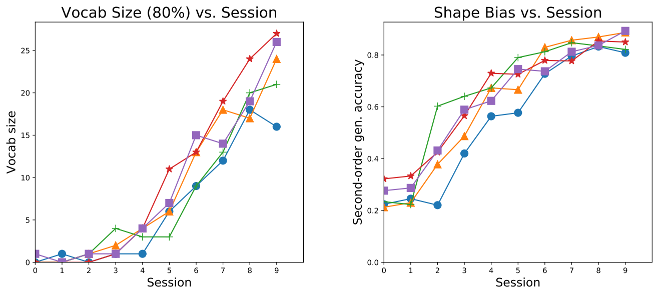


Figure 7: Learning curves for shape bias and vocabulary size of a CNN. TODO: update these plots, update caption.

The Onset of Vocabulary Acceleration

Our previous experiments confirm that simple neural networks can develop the shape bias when presented with realistic human training sets. It remains unclear, however, how this bias influences word learning. Gershkoff-Stowe & Smith (2004) showed that the development of the shape bias in infants is correlated with the onset of vocabulary accceleration; children who displayed a stronger shape bias in the study exhibited a higher rate of vocabulary acquisition. Eight children subjects participated in the study, beginning at ages 16 to 20 months. The vocabulary of each subject was monitored over the course of the study. Every three weeks, the shape bias level of each subject was evaluated in the lab using a test akin to the second-order generalization test of Smith et al. (2002). Authors of the study reported a number of interesting findings regarding the relationship between generalization performance and the number of nouns in a child's vocabulary. First, they found a correlation of 0.7 between generalization performance and the rate of change in the number of nouns possessed over the course of the study. TODO: discuss other findings of the study.

We replicated this experiment to the best of our ability using a CNN and our RGB image data. We trained a series of eight networks using 50 categories and 20 exemplars of each category. Fig. 7 shows plots that miror Figures 5 & 6 of Gershkoff-Stowe & Smith (2004). TODO: update this explaination. TODO: finish vocabulary acceleration experiments.

Discussion

Our experiments in this paper confirm that neural networks are capabable of developing the shape bias from as few as three examples of each object category. The development of this bias is known to improve the rate of word learning in human children, a phenomenon that is mirrored in our networks. One implication of this finding is that it may be possible to train large-scale image recognition models more efficiently after initializing these models with shape bias training. In future work, we hope to investigate this hypothesis with ImageNet CNN models using an initialization framework designed after the experiments in this paper.

Philosophers of science have discussed the importance of prior background knowledge for centuries (Tenenbaum et al., 2011). When the data provided to a learner is scarce, these priors help fill in the gap by constraining the space of models that the learner must consider. A new theory suggests that this phenomenon can be quantified from the perspective of information theory (Mattingly et al., 2017). Authors of the report show that, given an appropriate choice of prior over model space, a learner can maximize the amount of information extracted from limited data. They formalize this claim using the mutual information between the observed data and the parameters of the model. This measure quantifies the amount of information that can be learned about the model by measuring the data, or equivalently, the information about the data that can be encoded in the model. The information theory framework explains why structured model priors are helpful when data is scarce. The case of structured Bayesian priors is contrasted against that of a flat prior (i.e. no prior), with the former showing clear advantages.

TODO: final paragraph. What is left to discuss?

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.

Brodatz, P. (1966). *Textures: a photographic album for artists and designers*. New York, NY: Dover Publications.

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.

Dubuisson, M., & Jain, A. K. (1994). A modified hausdorff distance for object matching. In *Proceedings of the international conference on pattern recognition* (pp. 566–568).

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.

Mattingly, H. H., Transtrum, M. K., Abbott, M. C., & Machta, B. B. (2017). Rational ignorance: simpler models learn more information from finite data. *arXiv:1705.01166*.

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.