Learning Inductive Biases with Simple Neural Networks

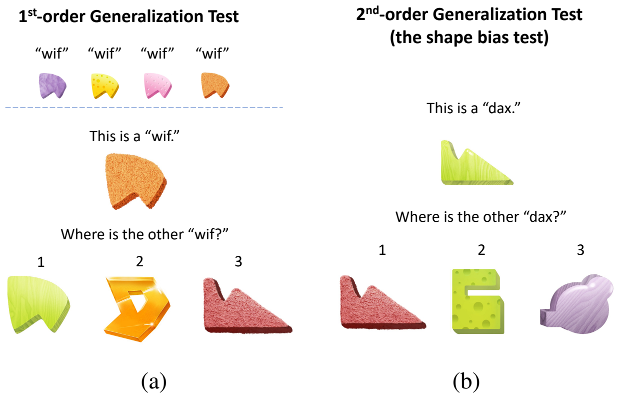
Reuben Feinman (reuben.feinman@nyu.edu)
Center for Neural Science
New York University

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 unrealistically large quantities of training data, and the conditions required for these biases to develop are not well understood. Moreover, it is unclear how the learning dynamics of these networks relate to developmental processes in childhood. 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 4 object categories. 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; word learning


1st-order Generalization Test
"wif"
"wif"
"wif"
"wif"
This is a "wif."
Where is the other "wif?"
1
2
3
2nd-order Generalization Test
(the shape bias test)
This is a "dax."
Where is the other "dax?"
1
2
3

(a) (b)

Figure 1: Shape bias generalization tests. The 1st-order test, shown in (a), assesses if a child has learned to generalize a familiar object name to a novel exemplar according to shape. This is the first step of shape bias development. The 2nd-order test, shown in (b), assesses if the child has learned to generalize a novel name to a novel exemplar by shape, the second and final step of shape bias development.

Introduction

Humans possess the remarkable ability to learn a new concept from seeing just a few examples. A child can learn the meaning of a new word such as "fork" after observing only one or a handful of different forks (Bloom, 2000). In contrast, state-of-the-art artificial learning systems use hundreds or thousands of examples per class 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 cognitive and neural 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" (Michalski et al., 2013), help the learner make inferences that go far beyond the observed data. For example, 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 novel name to new examples (Fig. 1b) (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 improve the efficiency of future learning (Harlow, 1949; Smith et al., 2002; Dewar & Xu, 2010).

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). However, it is currently difficult to apply HBMs to the type of high-dimensional visual and auditory stimuli that children receive. In some cases, HBMs and related approaches have been applied successfully to raw high-dimensional data (Lake et al., 2015), but only with the help of domain-specific knowledge and engineering. In contrast, current neural networks can learn effectively from many forms of raw data (LeCun et al., 2015), potentially providing the bridge between controlled simulations with synthetic data (e.g., Colunga & Smith, 2005) and large-scale real-

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

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 recurrent neural network, trained with sufficient experience via Hebbian learning, can acquire a shape bias for solid objects and a material bias for non-solid objects. These simulations demonstrate that neural networks can form different expectations for different kinds of objects, but they raise many new questions regarding the conditions required for neural networks to develop these types of biases. For example, Colunga & Smith (2005) used highly simplified bit-vector data, and it is unclear whether or not their findings generalize to more complex or realistic stimuli. Furthermore, the authors did not systematically vary the structure of the networks' experience, in terms of the number of categories and the number of examples, and thus we do not know the exact conditions in which biases arise and whether they can compete with the strong sample efficiency of HBMs (Kemp et al., 2007). In a recent study, building on advances in deep learning for object recognition, Ritter et al. (2017) found that performance-optimized deep neural networks (DNNs) develop the shape bias when trained on the popular ImageNet object recognition dataset consisting of raw naturalistic images. These results highlight an exciting possible connection between large-scale DNNs and developmental psychology, though many questions still remain. ImageNet–which contains about 1200 labeled examples of 1000 different object categories–is a poor proxy for the experience of a developing child, who typically develops a shape bias with no more than 50 or 100 object words in her vocabulary (Gershkoff-Stowe & Smith, 2004). Whether or not these networks can acquire the shape bias with more appropriate training sets is unclear. Furthermore, although 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 networks using artificial object stimuli that allow us to systematically vary the quality and form of the experience provided. Specifically, we use an experimental paradigm that is inspired by a developmental study performed with human children (Smith et al., 2002) to train and evaluate our networks, although the exact size of the learning sets differ in our setting. Beginning with simple bit-vector data akin to Colunga & Smith (2005), we systematically vary the number of categories and the number of examples in the training set, recording generalization performance of our networks on the 1$^{\text{st}}$- and 2$^{\text{nd}}$-order generalization tests of Smith et al. (2002) (Fig. 1) at each pairing. Parallel experiments are then performed with raw image data, where each image consists of a 2D object with a particular shape, color and texture. For both the bit-vector and image stimuli, we analyze the perceptual similarity scores of our corresponding networks as a function of physical stimulus distance along shape and color dimensions, gauging the parametric sensitivities to these attributes. In a final set of experiments, we studied the dynamics of learning-to-learn by analyzing the relationship between shape bias acquisition and the rate of word learning, mirroring an analogous study from developmental psychology (Gershkoff-Stowe & Smith, 2004).

Experiments

We set out to train neural networks with a learning paradigm used to guide toddlers to the shape bias (Smith et al., 2002). In this paradigm, the learner acquires new words that are organized exclusively by shape, such that different instances of the same category have the same shape, but contrast sharply in color and material. This is reflective of the fact that a child's early noun vocabulary consists predominantly of shape-based categories (Samuelson & Smith, 1999), although not with the same purity as provided in the shape bias training. As in previous computational modeling work (Kemp et al., 2007; Colunga & Smith, 2005), we focus on purified training with shape-based categories as administered in Smith et al. (2002), since it provides a controlled and specific test of the artificial learner's ability to make higher-order generalizations across varying quantities of training experience.

In Smith et al. (2002), 17-month-old children were taught four new object words ("wif", "dax", "zup", etc.) over the course of seven 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 novel objects were organized strictly by shape (two "wifs" have the same shape with differing colors and materials). During the weekly sessions, children played with each object while an adult announced the name of the object repeatedly. By the end of the study, the child participants had acquired a shape bias–i.e., they had formed the inductive bias that a novel name should be generalized by shape as opposed to color or material. A control group of children who did not partake in the play sessions did not form the same inductive bias.

We use the training paradigm of Smith et al. (2002) to study inductive bias learning in neural networks with artificial object datasets. We first perform our computational experiments with abstract bit-vector stimuli, followed by experiments with raw image data. The details of these two data formats and their corresponding network architectures are described in the succeeding two sections. Each constructed object is assigned a shape, color and texture value. We train simple neural networks to label objects with category names based on shape. To understand the necessary conditions for successful inductive bias learning, training is performed for various datasets with differing sizes and structures, varying both the number of categories and the number of examples of each category

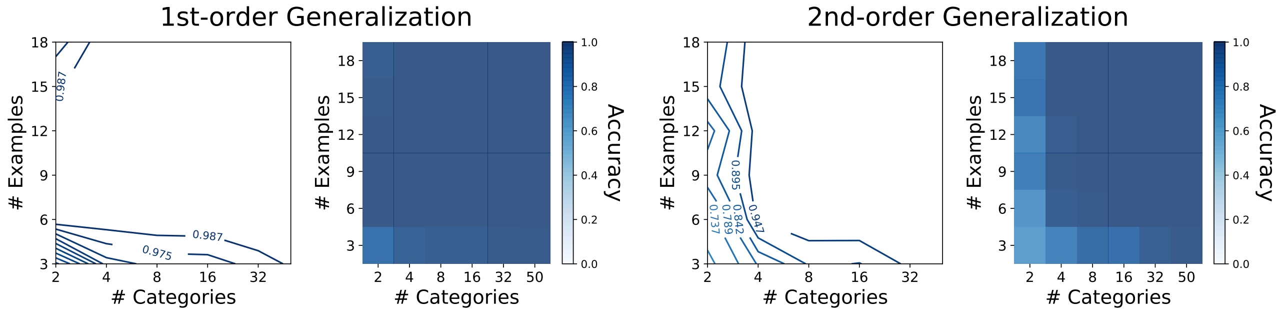


Figure 2: MLP generalization results for explicit shape bias training with various training set sizes. The number of categories and number of examples per category provided to the network are shown on the x and y axes, respectively. Plots show accuracy over 1000 trials of the specified generalization test, averaged from 10 training runs. The same data is shown in both contour and heatmap format.

provided to the network. The number of colors and textures are selected to match the number of categories in each case. 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), depicted in Fig. 1.

1st-order generalization test. For this test, toddlers are first presented with an exemplar object that they have seen and played with during training ("wif" in Fig. 1a). Then, they are presented with three test objects that they have not seen: one that matches the exemplar in shape (item 1 in Fig. 1a), one that matches in color (item 2), and one that matches in texture (item 3). For each potential match, the other two stimulus attributes are novel (untrained). The toddlers are asked to select which of the three test objects share the same name as the exemplar. Performance is measured as the fraction of trials in which the child selects the correct object, i.e. the shape match. Smith et al. (2002) propose that children who display the 1st-order generalization capability have taken the first step in the development of the shape bias: they have learned to attend to shape when identifying examples of familiar object categories. To simulate this test, we create an evaluation set containing groupings of four sample objects: an exemplar, a shape match, a color match, and a texture match. The activations of our network's last hidden layer are obtained in response to each object. We then evaluate the cosine similarity[2] between the activations of the exemplar and each test object to determine which object the network perceives to be most similar. Accuracy is defined as the fraction of groupings for which the correct (shape-similar) object is chosen.

2nd-order generalization test. For this test, toddlers are first presented with an exemplar object that has a novel label (e.g., "teema") as well as a novel shape, color and texture. From there, the trial proceeds similarly to those of the 1st-order: a shape match, color match and texture match are presented, and the child must select which test object she believes to share a name with the exemplar. All shapes, colors and textures are novel to the child in this test. Smith et al. (2002) propose that children who display the 2nd-order generalization capability have taken one step beyond the 1st-order generalization in the development of the shape bias. These children have not only learned to categorize a handful of object categories by shape, but they have induced that shape is a useful feature in general when categorizing objects (a form of "one-shot learning"). Learning the shape bias, as it was shown in Smith et al. (2002), is useful for vocabulary development. We simulate the 2nd-order test with artificial object stimuli similarly to the 1st-order case, again using last hidden layer features to evaluate network similarity scores.

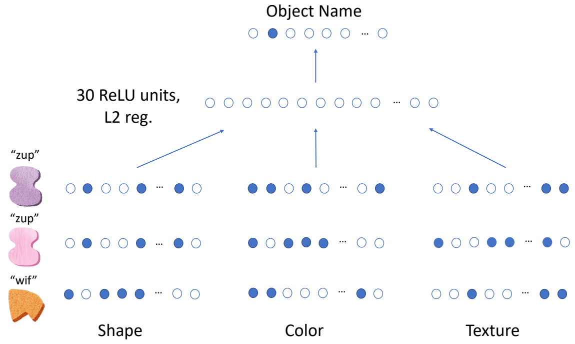


Figure 3: Multilayer perceptron architecture. Shape, color and texture attribute vectors are concatenated and fed to a 30-unit hidden layer, followed by a classification layer. 3 example input objects are shown (only one is presented at a time to the network).

In all simulations, we record accuracy over 1000 simulated test trials as the performance metric for each generalization.

Experiment 1: Multilayer perceptron trained on synthetic objects

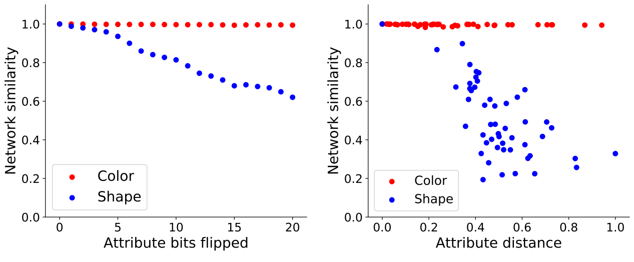
Our first experiment aims to study inductive bias learning in its purest form, using synthetic stimuli that can be exactly balanced and controlled. Objects are abstract binary patterns, divided into three input pools of 20 binary units each (representing the shape, color and texture of the objects; see Fig. 3). We vary the number of categories and number of examples per category in the training set. For datasets with N categories and K examples, we randomly generate N shape patterns, N color patterns, and N texture patterns. For all three

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

attributes, each pattern is replicated K times, ensuring equal entropy across the three. The shape patterns are then aligned with object labels, and the other two attributes are permuted randomly to create the dataset. A holdout set of shapes, colors and textures is retained for the generalization tests.

We train a multilayer perceptron (MLP) to name objects, as shown in Fig. 3. The network has an input layer of 60 units, a hidden layer of 30 rectified linear units (ReLUs) with L2 regularization, and a softmax output layer to classify the object by name. The softmax layer has N units (one for each label). We train the network for 200 epochs using negative log-likelihood loss, RMSProp, and batch size min(32, $\frac{N*K}{5}$).

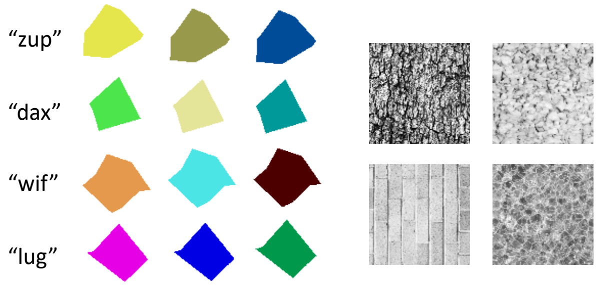
Results. Initially, as would be expected given the data format, shape is treated the same as other attributes. In the 2$^{\text{nd}}$-order generalization test, a randomly initialized network selects test objects with the following ratios, on average (50 trials): shape 35%, color 33% and texture 32%. We then trained the network with various dataset sizes. Results for the 1$^{\text{st}}$- and 2$^{\text{nd}}$-order generalizations are shown in Fig. 2, where each result is an average over 10 networks with different random seeds. We note that acquisition of the 1$^{\text{st}}$-order generalization requires less data than that of the 2$^{\text{nd}}$-order, as predicted by the 2-step hypothesis (Smith et al., 2002). Success in the 1$^{\text{st}}$-order test indicates that the network is learning successfully and generalizing to new examples of the training classes. Networks that have a shape bias score of 0.7 or higher on the 2$^{\text{nd}}$-order test have a substantial inductive bias, and the MLP reaches this threshold at the following points: N=2 & K=6 (accuracy 0.71) and N=4 & K=3 (accuracy 0.80). These results reproduce the general pattern of the Hierarchical Bayesian Model (HBM) in Kemp et al. (2007) and toddlers in Smith et al. (2002), who neared the 0.7 shape bias threshold with N=4 & K=2 (although the toddlers also received external experience). In contrast, Colunga & Smith (2005) used N=10 & K=100 to obtain the shape bias in their networks, using similar abstract patterns. Although HBMs are often noted for their data efficiency, in this case, the neural network was competitive for making 2$^{\text{nd}}$-order generalizations from limited data.



(a) MLP

(b) CNN

Figure 4: Perceptual similarity as a function of physical stimulus distance. A test stimuli is systematically altered along its shape or color dimension. Network similarity scores are computed between the original stimuli and its altered counterpart in each case, using the features of the last hidden layer.



(a)

(b)

Figure 5: Training stimuli for Experiment 2. (a) shows novel objects with various shapes and colors (the first three input channels). (b) shows a few examples of textures that might be found in the 4th input channel.

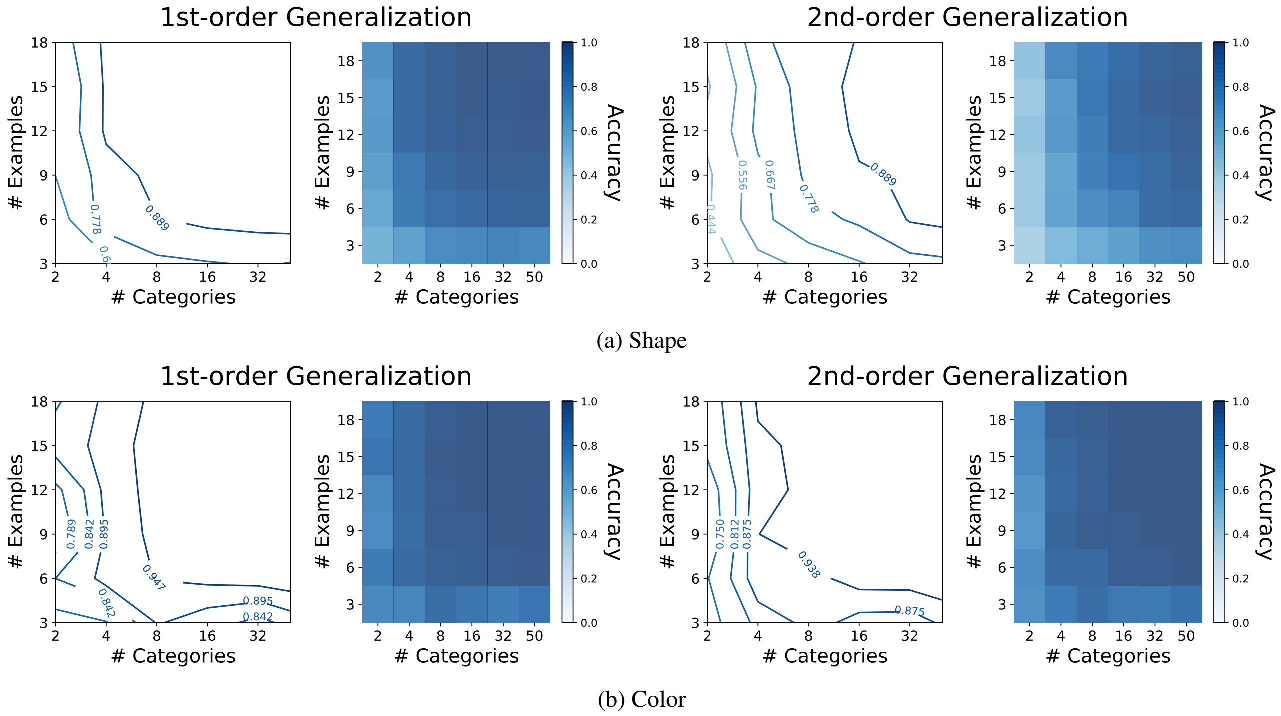
As another way of demonstrating the learned sensitivity to shape, we perform parametric manipulations of the stimuli. Using an MLP trained with N=4 & K=6, a training set that leads to a strong 2$^{\text{nd}}$-order score of 0.96, we probe the shape bias by selecting a novel test stimuli and systematically flipping b bits of the shape input pattern, recording the network similarity between the modified stimulus and the original for each b. For comparison, a similar test is also performed with color. Results are shown in Fig. 4a. Clearly, the network is far more sensitive to changes in shape than changes in color.

Experiment 2: Convolutional network trained on synthetic objects

Our first experiment used highly simplified training stimuli for maximal experimental control. One strength of modern neural network architectures is that they can learn effectively from data in raw and complex forms, a fact we take advantage of in developing Experiment 2. Here we ask whether similar learning-to-learn results can be achieved using synthetic object stimuli encoded as raw images. This setup presents a more challenging learning problem for the neural network, in terms of making both 1$^{\text{st}}$- and 2$^{\text{nd}}$-order generalizations, since understanding shape requires making abstractions that go substantially beyond separating a pool of input units that directly encode the attribute, as in Experiment 1.

The stimuli are constructed as follows. Each object is a 2D shape of a specified color placed over white background (200x200). Texture is represented in a fourth image channel, independent of RGB space.[3] Some examples of our objects are shown in Fig. 5. Object shapes are polygons of random order (uniform 3-10) and randomly sampled vertices, with preference given to points near image boundaries in order to ensure visible-sized objects. Colors are generated to span the

[3] In the experiments of Smith et al. (2002), children physically touch each object they are presented in addition to observing it. Although materials like plastic and styrofoam are 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 is logical to provide these signals along independent axes to a computational model.



(a) Shape

(b) Color

Figure 6: CNN generalization results. (a) shows results for explicit shape bias training, as discussed in Experiments. For comparison, (b) shows results for a network trained to label objects with category names based on color. In this case, the generalization tests evaluate the fraction of times that the color match is selected. Results in each grid show the average of 10 training runs.

RGB vector space with even separation. We use black and white textures from the Brodatz database (Brodatz, 1966) for our texture categories. A holdout set of shapes, colors and textures is again retained for testing.

We train a multi-layer convolutional neural network (CNN) (LeCun et al., 2015) consisting of two convolution layers with five feature maps, each followed by a max pooling layer. A depiction of this architecture is shown in Fig. 7. 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. Both the convolution layers and the fully-connected layer use L2 regularization, the latter also with dropout=0.5. Each object is randomly shifted around image space by a small offset (train and test alike). Training details mimic the MLP, but with 400 epochs.

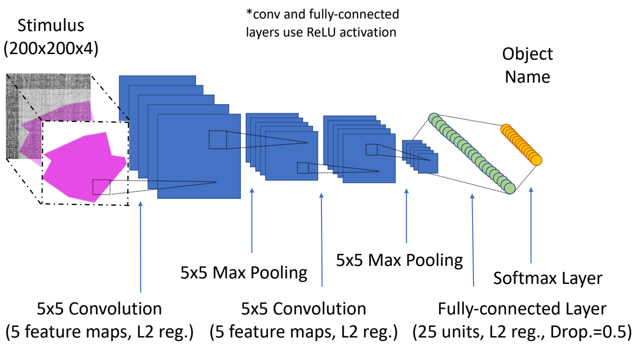
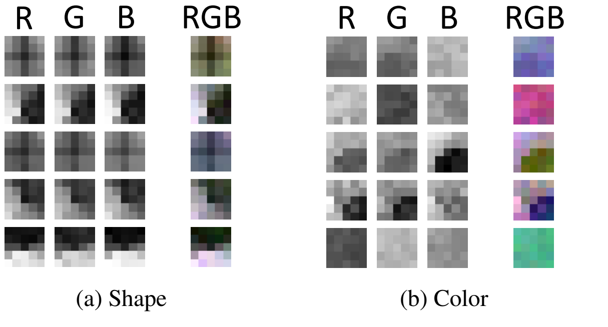

Stimulus
(200x200x4)
*conv and fully-connected
layers use ReLU activation
Object
Name
5x5 Max Pooling
5x5 Max Pooling
Softmax Layer
5x5 Convolution
(5 feature maps, L2 reg.)
5x5 Convolution
(5 feature maps, L2 reg.)
Fully-connected Layer
(25 units, L2 reg., Drop.=0.5)

Figure 7: Convolutional network architecture. The network receives 4D image stimuli and is trained to label the object in the image with a category name that is based on shape.

Results. The randomly initialized network makes 2nd-order selections with the following ratios: shape 38%, color 42% and texture 20%. We trained the network using varying dataset sizes, as with our MLP. Results are shown in Fig. 6a. Similarly to the MLP, we see that acquisition of the 1st-order generalization requires less data than that of the 2nd-order, supporting the notion that learning the training classes is a simpler task than forming higher-order generalizations. Using the same shape bias threshold of 0.7 2nd-order score, we find a number of important transition points: N=32 & K=3 (accuracy 0.74), N=8 & K=6 (accuracy 0.75), and N=4 & K=12 (accuracy 0.70). The CNN is thus capable of learning a shape bias from as few as 6 examples of 8 categories, a significant feat given the scale of the input. Notably, the network is able to learn this bias with much fewer data than Colunga & Smith (2005) using a data form that is significantly more complex. The CNN of Ritter et al. (2017) used roughly N=1000 & K=1200, and developed a shape bias of 0.68 on a shape and color-only task. A key takeaway from our results is that, with concentrated training effort, it is possible to learn this bias from much less data using high-dimensional color images.

As in Experiment 1, we parametrically manipulate the stimuli in order to analyze the network's sensitivity to changes along different stimulus dimensions, using a CNN trained with N=30 & K=10. This network achieves a strong 2nd-order score of 0.91. Distance in shape space is quantified as the Modified Hausdorff Distance (Dubuisson & Jain, 1994) between the shape pair. In color space, physical distance is quantified using the cosine distance of the RGB vector pair. Beginning with an exemplar object stimuli, we sample 50 secondary shapes and order them by their distance from the exemplar. We then modify the shape of the exemplar parametrically by stepping along this list from near to far and selecting the new shape, recording network similarities between the original and modified versions in each case. A mirroring experiment is then performed with color; in each case, only one attribute is altered at a time. Results are shown in Fig. 4b. As with the MLP, our CNN's selection preferences show a clear parametric dependency on shape, and a much weaker dependency on color.

For the sake of comparison, we also trained our CNN to label objects with names organized by color. Our goal was to compare the required sample complexity for color bias training with that of shape bias training, and to evaluate whether color bias development follows a similar 2-step process. All dataset parameters mirrored those of shape training, except that the object labels were aligned with the color attribute of each training image. Performance on the generalization tests was measured as the fraction of trials for which the network selects the color match. Results for CNN color bias training are shown in Fig. 6b. Notably, the color-trained CNN requires a smaller sample complexity to achieve 0.7 accuracy on the 2nd-order test, reaching a score of 0.73 with N=2 & K=3. Furthermore, this network does not appear to follow the 2-step process of bias development; results for 1st- and 2nd-order generalizations look near-identical to one another. In order to identify a stimulus as a member of a particular color category, the network needs only to find a single pixel of that color, a task that is much simpler than representing and identifying shape. Representing color requires a simple 3D space. By learning to isolate and preserve this space in the hidden layers, the model can easily generalize to novel colors, hence the early 2nd-order results. We inspected the learned representations of both a shape-trained and a color-trained CNN, trained with N=50 & K=18, by visualizing the first-layer convolution filters of each network (Fig. 8). As we would expect, filters of the shape-trained CNN look identical across R, G and B channels, as this network needs no sensitivity to color. In contrast, filters of the color-trained CNN vary across channels, indicating that the network has learned a selectivity for color.



(a) Shape (b) Color

Figure 8: Visualizing RGB channels of learned first-layer convolution filters. (a) shows the filters of our CNN trained with explicit shape bias training (N=50 & K=18). Each row corresponds to 1 of the 5 filters. The first 3 channels are shown in the 'R', 'G' and 'B' columns, respectively. These 3 channels are shown together in a 4th column, labeled 'RGB'. (b) shows mirroring filters for our CNN trained to label objects with category names based on color. In both (a) and (b), only channels 1-3 of the 4 are shown.

Experiment 3: The onset of vocabulary acceleration

Our previous experiments confirm that simple neural networks can develop the shape bias from a relatively small number of categories and examples. It remains unclear, however, how the dynamics of bias acquisition relate to the dynamics of word learning. Gershkoff-Stowe & Smith (2004) showed that the development of the shape bias in toddlers predicts the onset of vocabulary acceleration during early word learning, a phase that begins at ages 16-20 months. Studying 8 children during regular lab sessions at 3-week intervals, the authors found that increasing attention to shape was correlated with increasing rate of vocabulary acquisition in participants. Fig. 9a shows the individual growth curves of vocabulary size and shape response for each child. The former variable is measured as the cumulative number of nouns in the child's vocabulary, and the latter is measured as the cumulative number of times that the child has selected the shape match in a shape bias task. Although the vocabulary curve shows cumulative nouns in whole, the authors also recorded cumulative "count nouns" for each participant, a particular subset of nouns that is well organized by shape. We focus on the statistics reported for count nouns, as this subset more directly taps into the type of vocabulary that is influenced by the shape bias.

Authors of the study reported three interesting correlations: **1)** a correlation of 0.75 between increase in cumulative shape choices and increase in cumulative count nouns across sessions for an individual participant, averaged over participants ($p < 0.05$),[4] **2)** a correlation of 0.81 between the average increase in shape choices over the experiment and average increase in count nouns, computed across participants ($p < 0.05$), and **3)** a correlation of 0.85 between the index of the first session in which a child shows a "systematic" shape bias and that of the first session in which she shows a "substantive" increase in count nouns, computed across participants ($p < 0.01$).

Methods. Inspired by this study, we train a CNN using our raw image data with the goal of evaluating related correlation metrics for our networks. The participants of Gershkoff-Stowe & Smith (2004) were not explicitly trained for the shape bias as done in Smith et al. (2002); they received natural experience in a home setting, which may have included some words organized by attributes other than shape. Therefore, we design a new learning framework for our CNN in

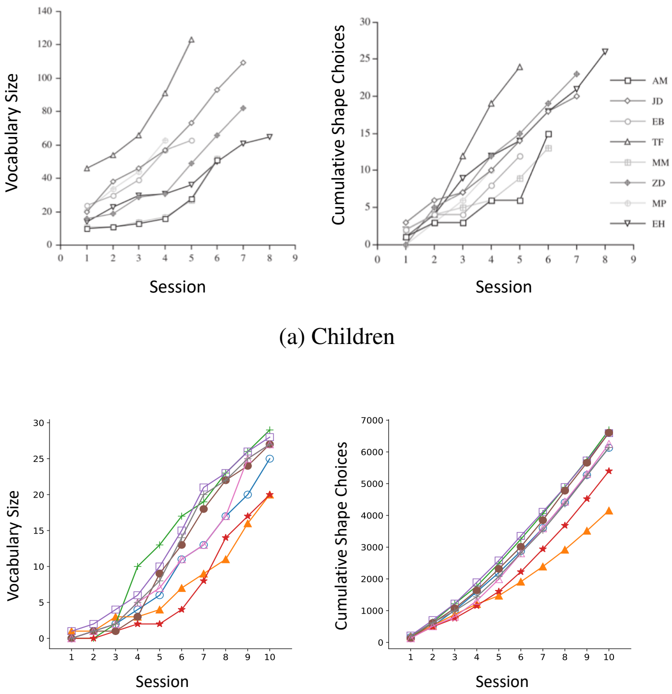
[4] p-values are one-tailed.

this setting. We train our CNN to simultaneously label the object's name, which correlates with shape, as well as its color and texture names. The number of categories along each label dimension and the loss weight assigned to that dimension are determined according to the natural statistics of the early human lexicon (Samuelson & Smith, 1999).[5] The chosen ratios are as follows: 60-20-20 shape-color-texture names (36, 12 and 12 categories, respectively). 10 examples of each shape are used, and colors and textures are assigned at random to each stimuli from their 12 categories.

We keep a cumulative count of the number of count nouns in the network's vocabulary, defined as the number of shape categories for which the network has achieved 80% or greater accuracy on the training set. We also keep a cumulative count of shape choices the network makes in a 500-trial 2nd-order test. This process is repeated with 20 networks, using a different random seed for each network.

Results. We inspect the "early" word learning period for our networks, defined as the period in which the average vocabulary size across the 20 networks is less than or equal to $2/3$ the total number of count nouns. Beyond this period, which we find to include the first 30 training epochs, the network's learning begins to flatten. We divide this period into 10 "sessions," evenly spaced by 3 epochs. The learning curves of our networks are shown in Fig. 9b. We compute correlation metrics for our networks that are analogous to those of the child study. Looking at increases across the sessions of a single network (metric **1**), we find an average correlation of 0.53 between increase in cumulative shape choices and increase in cumulative count nouns, with 10 networks showing $p < 0.05$. Further, looking at average increases across the entire 10-session period for each network (metric **2**), we find a correlation of 0.76 ($p < 0.001$) across the 20 networks.

To compute an analogous measure for metric **3**, we need to define our thresholds for "systematic" shape bias and "substantive" increase in vocabulary. Gershkoff-Stowe & Smith (2004) define the former as the first session in which a child exhibits a performance on the shape bias test that, if the child were selecting matches in this test at random, would only occur with probability 0.1. In our framework, we have 1500 test trials during each session. Using a binomial test, we can reject the null hypothesis that the network is responding randomly with $p < 0.1$ given 523 or more shape choices out of the 1500 trials. The authors define a "substantive" increase in vocabulary size as the first session that vocabulary size increases by 10 from the previous session. Since they choose threshold 10 for a maximum vocabulary of 100 words, we use threshold 4 (i.e. 3.6) for our 36-word vocabulary. With these thresholds, we find a correlation of 0.52 ($p < 0.015$).



(a) Children

(b) CNN

Figure 9: Learning curves for shape bias and vocabulary. (a) shows the learning curves of the 8 children participants from Gershkoff-Stowe & Smith (2004). Participants were studied over the course of 5-8 lab sessions. Curves are shown for vocabulary size (left) and cumulative shape choices (right). Here, vocabulary includes all noun types. (b) shows analogous plots for our CNN networks. 8 networks are shown, randomly sampled from the total 20 for the sake of visibility. Here, vocabulary is measured only for shape-based object names.

These analyses confirm that the dynamics of shape bias acquisition and early word learning show a considerable dependency on one another in our CNN models, a phenomenon that is mirrored in the early word learning of human children.

General Discussion

Using a set of controlled synthetic experiments, our work in this paper provides novel insights about the environmental conditions that enable learning-to-learn in neural networks. Building on the work of Colunga & Smith (2005), we show that simple neural networks can learn the shape bias from stimuli presented as abstract patterns with as few as 3 examples of 4 categories. Thus, these networks approach both HBMs and human children in the sample complexity required for bias development. Expanding on Ritter et al. (2017), we show that simple CNN architectures, trained with high-dimensional color images, are capable of learning the shape bias with as few as 6 examples of 8 object categories. Not all stimulus attributes are the same, however; our experiments with color training show that the learning dynamics of CNNs vary depending on the attribute they are trained to attend to. This result highlights the importance of scaling up simula-

[5] Children are taught object names, color names and material names independently. Loss weighting provides a good analog to this with a framework suitable to CNN training. Assigning a weight of 0.6 to object name labeling mirrors presenting this type of name 60% of the time in training. Similarly, a weight of 0.2 to color name labeling mirrors 20% presentation, etc.

tions to more complex data, where the presentations of different stimulus attributes can differ in physical form. Finally, we present novel results showing how a trained network's sensitivity to shape varies parametrically with the input.

In a very recent study, Hill et al. (2017) trained a neural network agent to navigate around a virtual 3D word and collect objects according to name-based language commands, using simplified artificial object stimuli similar to our own. The authors draw inspiration from studies with human children, and their results are noteworthy; however, the goal of our work differs in important ways. The agent in this experiment receives visual and language inputs together in conjunction, and must output navigation decisions. Thus, the network is asked to learn a variety of tasks simultaneously–namely, visual perception, language comprehension and navigation. The learning curves of the paper therefore reflect a form of multifaceted learning. Our primary interest in this paper is to study the precise quantity of data required for a neural network to learn inductive biases. Our framework is designed to investigate this question in isolation, with minimal interference from external factors.

The development of the shape bias in human children is known to correlate with improved word learning, 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-scale DNNs, using an initialization framework designed with the intuitions garnered here.

Acknowledgements

This research was supported by a Henry M. McCracken fellowship. 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.

Hill, F., Hermann, K. M., Blunsom, P., & Clark, S. (2017). Grounded language learning in a 3d simulated world. *arXiv preprint arXiv:1710.09867*.

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.

Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (2013). *Machine learning: an artificial intelligence approach*. Berlin, Germany: Springer Science and Business Media.

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).

Samuelson, L. K., & Smith, L. B. (1999). Early noun vocabularies: do ontology, category structure and syntax correspond? *Cognition*, *73*(1), 1–33.

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* (pp. 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.