**CogSci 2018 paper**

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 far beyond the observed data. A recent study exposed that performance-optimized deep neural networks develop the shape bias, an inductive bias possessed by children that plays an important role in early word learning. However, these models use unrealistic training data, and it remains unclear whether they develop their biases in the same way as humans. We investigate the development and influence of the shape bias in neural networks using artificial toy data of both categorical and image form that allows us to closely mimic the train and test datasets that were presented to human children during developmental studies. We find that simple neural network models display the shape bias after seeing as few as 3 examples of each concept, and that these biases tend to grow 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 human children.

1. 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 visual concept “fork” after observing only one or a handful of forks (Bloom 2000). In contrast, state-of-the-art artificial learning systems use many thousands of examples to learn simple visual concepts (Krizhevsky et al. 2012, Szegedy et al. 2015). Although the representations of these systems show parallels to biological vision (Yamins et al. 2014), biological learning systems use far fewer labelled examples. 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 concept learning.

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 isn’t 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).

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) are a popular approach whereby probabilistic inference occurs at multiple levels simultaneously, allowing the model to learn a prior that is used when processing subsequent data (Kemp et al. 2007, Salakhutdinov et al. 2012). Although these models show promise, they fall short of human capabilities in critical ways. One the one hand, it is difficult to apply HBMs to the type of raw, high-dimensional sensory data that children receive, such as RGB pixels or audio waves. On the other hand, these models must be hand-designed for a particular task of interest, e.g. handwritten character learning or grammar learning. In contrast, neural network models perform well with raw data formats across a range of domains including image processing, speech recognition, drug discovery and genomics (LeCun et al. 2015), relying in each case on a canonical architecture with multiple layers of nonlinear processing. 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 models operate on bit-vector data, and it remains unknown whether a similar feat can be performed with RGB images. Furthermore, the authors do not systematically very the structure of the training set, so we don’t know the exact conditions in which biases arise, nor whether current models are sufficient to explain it.

A recent study 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 RGB images (Ritter et al. 2017). These results highlight an exciting connection between DNNs and developmental psychology; however, a significant gap lingers between the two ontologies. 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 inductive 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).