**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 (DNNs) develop the shape bias, an inductive bias possessed by children that is critical to 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 DNNs using an experimental paradigm borrowed from developmental psychology. 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 just one or a handful of examples. A child that is learning her first few words can easily pick up the meaning of the visual concept “dog” after just one or two presentations (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, GoogleNet, etc.). Although these systems exhibit a few promising connections to biology (Yamins et al. 2014, Ritter et al. 2017), their training sets contain far more examples than any human being will see in a lifetime. Consequently, significant effort is ongoing to understand what neural and cognitive mechanisms enable efficient concept learning (Lake et al. 2017).

One prominent theory holds that humans use prior background knowledge in order to delimit the hypothesis space during induction (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 of the same shape are of the same category—when learning new object names (Fig. 1) (Landau et al. 1988). Similarly, children assume that object names are mutually exclusive, i.e. that once a name has been used for a certain object type it may not be reused for another (Markman & Wachtel 1988). Although some inductive biases may be innate, results from human and primate studies alike suggest that a handful of these biases can be learned, and that the acquisition of these biases can improve the efficiency of future learning (Harlow 1949, Smith et al. 2002).

**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 correct 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) are a popular approach whereby a Bayesian prior is first learned from data and then used for structured probabilistic inference in successive experiments (Kemp et al. 2007, Salakhutdinov et al. 2012). Although these models show promise, they rely on symbolic computation, and they fall short of human capabilities in critical ways. One the one hand, these models have difficulty with raw sensory data such as RGB pixels or audio waves. On the other hand, HBMs 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. Beyond HBMs, … TODO: discuss Colunga & Smith paper, discuss why we’re different.

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 (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. How the DNN shape bias varies as a function of the training set size, and whether or not these models can acquire the same bias with a training set comparable to humans, remains unknown. 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 mimic psychological 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 data sets, varying both the number of categories and the number of exemplars of each category. For each of our models, we investigate the parametric relationship between bias strength and attribute similarity 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. Beyond the generalization tests, we perform an additional experiment to 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). Need a sentence here about our results?