**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. Recently, it was discovered that performance-optimized deep neural networks (DNNs) develop an important inductive bias for word learning—the shape bias—which is also possessed by children. 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 inductive biases in DNNs using an experimental paradigm borrowed from developmental psychology. We find that simple neural network models can develop inductive biases from 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.

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 (cite AlexNet, 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. In 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 are learned, and that the acquisition of these biases can improve the efficiency of future learning (Smith et al. 2002, Harlow 1949).

Cognitive scientists have proposed a number of computational models to explain how inductive biases are acquired and harnessed for future learning (Kemp et al. 2007, Salakhutdinov et al. 2012). However, these models rely on conceptual-level symbolic description, and they fall short of human capabilities in critical ways. One the one hand, these models cannot operate on raw sensory data, 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 hierarchical architecture with multiple layers of nonlinear processing. TODO.