**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 available 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

Human learners…