Using Bayesian Inference to Understand Inductive Biases in Deep Neural Networks

Abstract

Human-like intelligence in machines has been of interest to philosophers and engineers for centuries. Today's deep learning allows computational models to obtain meaningful representations on high-dimensional inputs such as text and images. However, understanding these representations is challenging and why they work remains opaque. Our aim goes in two directions: 1) using computational foundations of human cognition to understand and to improve deep neural networks; 2) reverse-engineering the mind by using success in deep neural networks to understand human cognitive processes. In both directions, Bayesian inference serves as a bridge between the two fields. In the first goal, we present a Bayesian interpretation of the autoregressive objective under several distributional assumptions to show where the optimal content of language model embeddings can be identified. In the second goal, we show how variants of prior distributions in variational auto-encoders, a probabilistic generative model, explain how forming prototypes contributes to semi-supervised categorization.

Textbook

Bishop (2006) - Pattern Recognition and Machine Learning

Papers

Language Modeling

1. Elman (1990) - Finding Structure in Time
2. Bengio, et al. (2003) - A Neural Probabilistic Language Model
3. Blei, et al. (2003) - Latent Dirichlet Allocation
4. Radford, et al. (2018) - Improving Language Understanding by Generative Pre-Training
5. Andreas (2022) - Language Models as Agent Models
6. Cunningham, et al. (2023) - Sparse autoencoders find highly interpretable features in LLMs

Bayesian Deep Learning

1. MacKay (1995) - Probable networks and plausible predictions? A review of practical bayesian methods for supervised neural networks
2. Jordan, et al. (1999) - An Introduction to Variational Methods for Graphical Models
3. Griffiths, et al. (2006) - Bayesian models of cognition
4. Blei (2014) - Data analysis with latent variable models
5. Kingma & Welling (2014) - Auto-encoding variational Bayes
6. Gruver, et al. (2023) - Large Language Models Are Zero-Shot Time Series Forecasters

7. McCoy & Griffiths (2023) - Modeling rapid language learning by distilling Bayesian priors into artificial neural networks