A Modular Framework for Analyzing Theory of Mind Learning in Competitive Tasks

Joel Michelson
Vanderbilt University, Nashville, Tennessee, United States

Deepayan Sanyal
Vanderbilt University, Nashville, Tennessee, United States

Maithilee Kunda
Vanderbilt University, Nashville, Tennessee, United States

Abstract

A key challenge of theory of mind, or the ability to reason about others' mental states, is understanding the process by which others' perceptions influence their beliefs. While specific tasks—like competitive feeding—benchmark participants' ability to infer beliefs, it remains unclear how such capabilities can be learned. In this work, we introduce a modular framework that solves a computational, competitive-feeding-like game in which two agents compete. By systematically replacing modules of a successful rule-based framework with neural networks, we identify which capabilities can be learned from narrow sets of experiences, and which are critical for robust generalization. Using feature extraction techniques, we analyze how different architectures process task-relevant information. Finally, we describe and compare three novel approaches to improving generalization via first-person exposure to uncertainty: role reversal with the opponent, artificial observation masking, and synthesizing beliefs from conflicting information.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*