Inferring actions, intentions, and causal relations in a neural network

Andrew Saxe
Harvard University, Cambridge, MA, USA

Abstract: From a young age, we can select actions to achieve desired goals, infer the goals of other agents, and learn causal relations in our environment through social interactions. Crucially, these abilities are productive or generative: for instance, we can impute desires to others that we have never held ourselves. This capacity has been captured by the powerful Bayesian Theory of Mind formalism, but it remains to forge connections to the rich neural data around action selection, goal inference, and social causal learning. How can productive inference about actions and intentions arise within the neural circuitry of the brain? Using the recently-developed linearly solvable Markov decision process, we present a neural network model which permits a distributed representation of tasks. Such a representation allows the expression of infinite possibilities by combining a finite set of bases, enabling truly generative inference of actions, goals, and causal relations in a neural network framework.