Machine Theory of Mind

Björn Bebensee (2019–21343)
Topics in Artificial Intelligence

September 19, 2019

Theory of Mind (in humans) refers to the ability to reproduce the mental state of others and understand their intentions, goals, beliefs and desires which allows us to predict their future actions. This process works without any knowledge of the underlying structure: rather than knowing another's brain's physical activity, we understand their more abstract mental state. Rabinowitz et al. propose a *Machine Theory of Mind* which attempts to model the behaviour of an agent from observations in order to make predictions about its future actions. They believe that while difficult to achieve, building a machine ToM can benefit many fields. One such application that they propose is an intermediating system between an artificial agent and humans that may help interpretability of these agents. In this paper they focus on formulating the problem and demonstrate the viability of their approach to this problem through a serious of experiments.

Rabinowitz et al. consider building a Theory of Mind as a meta-learning problem in their work. In this setting, an observer gets access to behavioural traces which it then uses to learn to make predictions of the agent's future behaviour. In the training phase the observer should get better at making these predictions and learn a strong prior which it can then employ to make faster and better predictions. This is different from previous work in the field which mostly focused on hand-crafted models of agents rather than a meta-learning approach to the problem space.

As a means to achieve this goal the authors propose a Theory of Mind neural network (*ToMnet*). This ToMnet consists of three main components: a *character net* which parses observations yielding an embedding e_{char}, a *mental state net* which attempts to infer the agent's current mental state given past observations e_{char} yielding a mental state embedding e_{mental}, and finally a *prediction net* which leverages both the character and state embeddings to to predict the agent's future behaviour. This prediction net outputs next-step predictions of an agent $\hat{\pi}$ (its policy) as a vector of probabilities of future actions, a prediction whether any objects will be consumed by the end of the episode $\hat{c}$ as well as a prediction $\hat{SR}$ of the agent's trajectory in the current MDP (its successor representation).

While only focusing on POMDPs in grid-worlds, the experiments show the viability of this meta-learning approach. ToMnet learns a model which generalizes from agents during the training phase and then builds upon this prior to construct a model from observations of a specific agents online. These can be leveraged to make predictions of its future behaviour. However, as these experiments are limited and very simple in their form, this approach requires future research and it is not entirely clear from these experiments whether it will scale well to more complex domains.

References

[1] Rabinowitz, N. C., Perbet, F., Song, H. F., Zhang, C., Eslami, S. M., Botvinick, M. (2018). Machine theory of mind. *arXiv preprint arXiv:1802.07740*.