Review of
“Machine Theory of Mind”

Robin Khatri

December 2, 2018

This is a review of “Machine Theory of Mind” published in the proceedings of the 35th International Conference on Machine Learning, PMLR 80:4218-4227, 2018. The original paper is available at http://proceedings.mlr.press/v80/rabinowitz18a.html.

In this paper, the authors have trained an autonomous system that successfully learns how to model other agents in its world using only a small number of observations. The authors have referred to this system as a Theory of Mind Neural Network (ToMnet). Further, the authors have termed this theory as Machine Theory of Mind. The name is a derivative of Theory of Mind [1]. The authors have experimented with ToMnet in Partially Observable Markov Decision Processes (PMCDPs). ToMnet encounters agents in a gridworld; it learns about their behavior without the access to their parameters. The authors have included different species of agents. These different species of agents include random or stochastic agents, algorithmic agents and Deep Reinforcement Learning (RL) agents. Moreover, the authors show that the ToMnet that has been trained specifically on a distribution of agents, is capable of producting agent-specific theory of Mind estimations. The authors have considered both simple and complex experiments to test ToMnet. The authors have also tested ToMnet on classic Theory of Mind tasks such as Sally-Anne Test [2]. For this purpose, the authors created a variant of Sally-Anne Test in a gridworld.

Pros:

- + An interesting study about building a nueral network capable of doing tasks that are associated with the general Theory of Mind.
- + Both stochastic and learned agents are considered.
- + Every agent which ToMnet encounters has its own observation functions, reward functions and discount factors. This allows for simulating real world. In real world, humans differ in their thought process and their beliefs.
- + During experiments, observer (ToMnet) does not have access to parameters and functions associated with the agents it encounters.
- + Same ToMnet architecture system is capable to provide estimations of agent-specific Theory of Mind tasks.
- + The authors have shown that ToMnet passes such Theory of Mind Tasks such as knowing that others can have false beliefs about the world.

Cons:

- − Only gridworlds are used as environments. Generalisation suggested to 3-D visual environments is presented without evidence.
- − While structure of Gridworld is detailed in appendices, no source has been provided for definition of Gridworld. There has been lack of citations to provide literature on other terms such as RL agents.

- There has been a few instances of ambiguity in suggesting generalisation of ToMnet results.
- Some considerations and assumptions taken in building ToMnet are unexplained.

Overall, I very much liked this paper as it has provided a methodology in which autonomous modelling of behavior of others in possible. The paper has been successful in implementation of experiments presented in seminal work of Bayesian Theory of Mind [3] [4]. Adding to this, this paper also attempts to solve challenge of using meta-learning in developing a Theory of Mind applicable to autonomous systems. The understanding of this topic is important for improvement in human understanding of Artificial Intelligence. It also has several real world applications such as autonomous decision-making in multi-agent tasks.

My concerns regarding this paper are directly related to the list written above.
Notably,

1. There are some phrases and statements in this paper that need to be explained in more detail:

   a. In this paper, only single-agents Partially Observale Markov Decision Processes (POMDPs) are considered. The authors have mentioned that the results should also generalise for multi-agent Decision Processes. Is there any evidence in support of this generalisation?

   b. The authors have made an assumption that the agent does not carry its hidden states from one episode to another. Is it a necessary condition in developing a system like ToMnet? If not, what will be the behavior of ToMnet when there is a vilation of this assumption?

   c. The authors omitted mental net from ToMnet architecture during the training of ToMnet on random agents. Was it because mental net was not needed for this training? or was there some other consideration?

2. Questions regarding methodology and empirical results:

   a. Does the behavior of ToMnet change if we use a different methodology in training the Deep RL agents that ToMnet encounters?

   b. In inferring goal-oriented behavior of agents, the authors enrich the agent species by applying a very high move cost (0.5) to 20% of the agents. What are the motivations behind this specific cost?

   c. During the testing of ToMnet on the variant of Sally-Anne Test, agent policy changed when sawpping was in a view in 2 block radius. What is the reason behind it being 2 block radius? Is this an empirical evidence or there is some theoretical reason behind this?

3. Authors' identities

   Since this is a published paper, indentities of the author were already mentioned in the downloaded paper.

References

[1] D. Premack and G. Woodruff, "Does the chimpanzee have a theory of mind?," *Behavioral and Brain Sciences*, vol. 1, p. 515, dec.

[2] S. Baron-Cohen, A. M. Leslie, and U. Frith, "Does the autistic child have a "theory of mind" ?," *Cognition*, vol. 21, pp. 37–46, oct 1985.

[3] C. L. Baker, R. R. Saxe, and J. B. Tenenbaum, "Bayesian theory of mind: Modeling joint belief-desire attribution," in *In Proceedings of the Thirtieth Third Annual Conference of the Cognitive Science Society*, pp. 2469–2474, 2011.

[4] C. L. Baker, J. Jara-Ettinger, R. Saxe, and J. B. Tenenbaum, "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing," *Nature Human Behaviour*, vol. 1, mar 2017.