Incremental Learning of Mental Models for Behavior Understanding

Doctoral Consortium

Jan Pöppel
CITEC, Bielefeld University
Bielefeld, Germany
jpoeppel@techfak.uni-bielefeld.de

KEYWORDS

Theory of Mind; Action Understanding; Reasoning; Model Learning

ACM Reference Format:

Jan Pöppel. 2018. Incremental Learning of Mental Models for Behavior Understanding. In *Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), Stockholm, Sweden, July 10–15, 2018,* IFAAMAS, 3 pages.

1 MOTIVATION

Humans are inherently social beings. Starting from a very young age, they quickly become increasingly more competent at inferring other's intentions from their behavior [18]. This represents a crucial skill, as inferring other's intention allows cooperation without explicit verbal communication as well as preparing for the next likely actions of those around. What is even more impressive is that humans are capable of inferring reasons for seemingly suboptimal behavior, allowing them to collect much more information about the agent they are observing as well as the situation that agent is currently in. This is possible because they are assuming that other humans are rational agents, meaning that their actions should serve a purpose, it might just be that they cannot immediately understand this purpose. Children can use this additional information to learn more about the world around them [11].

This capability, which goes beyond mere goal or action recognition, can also benefit computational systems, by allowing them to anticipate their users' next action or intention, without requiring explicit communication, thus greatly improving human computer interaction and potential cooperation with robots [10].

Such a system requires a mental model of its users which is capable of explaining their behavior. Such a mental model of oneself or others has been named Theory of Mind (ToM) in human cognition and psychology [14]. There exist differing interpretations of how ToM works in humans, the most prominent views can be summarized as the "Theory-theory", which states that humans have a functional module performing these mental inferences for them, and the "Simulation Theory", which states that they use already existing modules for action planning and recognition to simulate the behavior and mental states of others [7].

For computational systems, recent work tends to lean slightly towards the Theory-Theory with the most prominent approach being the Bayesian Theory of Mind by Baker et al. [2] which views the problem of mentalizing as inverse planning in a Bayesian framework relating mental states and observed behavior. There have been several studies showing that this inverse planning matches well with human judgment in a number of differing situations ranging from goal or intention recognition (e.g. [6, 16]) including compositional desires [17] to plan inference [4] and even up to the inference of potentially false beliefs [2] and preferences [8].

The main problem with this approach is that "[the inverse planning problem is ill-posed and] requires strong prior knowledge of the structure and content of agents' mental states, and the ability to search over and evaluate a potentially very large space of possible mental state interpretations." [3] The prior knowledge requirement results in specific mental models tailored for specific scenarios, while the evaluation requirement usually limits the previously considered models to only a couple of different mental variables with a finite set of potential states. Diaconescu et al. [6] proposed the use of hierarchical Bayesian models in order to cover a greater range of relevant mental states, but even this requires careful prior design and inference becomes increasingly more difficult the more mental states are considered, making very large and complex model quickly intractable. This is due to the fact that in order to compute the normalization required for exact Bayesian inference, one needs to marginalize over all mental states. The computational complexity thus directly increases exponentially with the number of mental states.

Specifically designed models however have their own problems: Each of these models will only work well within the specific context it was designed for. A change in the environment, or even just in the task, might require a modification of the considered mental states, in order to still provide acceptable predictions. Summarizing, current computational systems for mentalizing as a way of behavior understanding are usually not general, adaptive or not efficient enough to be used in real-time human machine interactions.

As a result of this, I consider the following core question in my thesis which is supervised by apl. Prof. Dr.-Ing. Stefan Kopp:
How can an artificial agent incrementally learn to explain observed behavior in a satisfying manner and how can the learned model(s) be efficiently evaluated in real time?

I should note that I do not try to replicate the mechanism by which humans acquire these mentalizing skills although the developed methods may be inspired and tested against findings in human (developmental) cognition. Instead my goal would be for artificial systems to learn to understand observed behavior (usually performed by humans) *well enough* to be useful in a wide range of scenarios, including online assistance and interaction with users.

Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), M. Dastani, G. Sukthankar, E. André, S. Koenig (eds.), July 10–15, 2018, Stockholm, Sweden.

What should be considered as "well enough" will depend on the situation but is always made up of both accuracy as well as efficiency. In some situations, it would suffice to predict an agent's next actions, therefore *well enough* would be primarily defined by a low prediction error. When inferring the mental states of the observed agent, *well enough* should be measured by comparing to human judgments of the same observations or ratings of the inferred mental states. In more complex interactions, where the system needs to cooperate with the observed agent, *well enough* would be measured by both parties' average utility while real-time performance of the system would also be important. Importantly, by focusing on the expected utility, the system is allowed to make uncertain or even incorrect inferences regarding some of the mental states of the agent, as long as it still chooses suitable actions, an approach previously demonstrated in [15].

2 PREVIOUS WORK

So far I have followed the Bayesian Theory of Mind approach of inverse planning to infer underlying goals or intention from observed behavior. I have developed a similar 2D domain as the food truck scenario employed by Baker et al. [1] by creating a maze navigation task. By modifying the amount of information about the task or the environment available to the navigating agent, I induce different sources of uncertainty, thereby creating different scenarios. If the task is to reach a specific goal, I can for example show additional potential goal positions, which are only distinguishable from close distance, thus introducing uncertainty regarding the true goal position. A system does not need to consider a mental state representing the goal positions, when there are no distracting goal positions. However, in order to successfully explain exploration behavior in light of distracting goals, a more complex model is required.

In order to cope with multiple scenarios, each with different sources of uncertainty would require a model incorporating all of these uncertainties, which can quickly become quite complex from both the conceptual and computational perspective.

Within this environment, I proposed a system for dealing with a range of different scenarios, by considering and switching between a range of simpler explanation. This idea was inspired by the intuition, that humans will often employ simple explanations as long as they suffice, which is also supported by findings in human cognition literature [5, 12].

2.1 Model switching

In [13] I was able to show that specialized models, each only considering the mental states relevant for one given scenario, were able to perform well within their given context. However, their performance was unsurprisingly a lot worse when applied to other scenarios. A complex model, containing the mental states required in all situations performs better overall, but usually worse than the specialized models in their given contexts. The mentioned proposed system starts with the simplest model and only considers switching to another one when its predictive performance becomes too poor. So far I measured this performance by comparing the predicted behavior to the observed behavior.

I was able to show that even a very simple switching strategy outperforms both the specialized models as well as the complex model in all conditions, while being a lot more computationally efficient than the complex model. These results mirror the findings in cognitive science that humans appear to often employ heuristics (or biases) when making inferences (see e.g. [9]).

I further see these results as an indication, that artificial systems for behavior understanding can work very well by relying on simple explanations, or heuristics.

3 FUTURE WORK

So far, I have only considered *incremental learning* in the sense of incrementally updating my current model, both within each of the specialized models and by switching which model to consider. A key aspect of my thesis is to also develop a framework which can incrementally modify existing models and ideally also learn suitable new models for a given scenario through observation of behavior as well as interaction.

Obviously, this is a vastly underspecified problem which will likely be impossible to solve in general. Determining which assumptions are required for the system to successfully learn such models will also be part of my ongoing work.

A first step lies in the modification and expansion of previously known mental models, i.e. within already known mental structures. This includes learning about previously unknown intentions through observations. Towards this goal, I am currently working around the rational agent assumption. By assuming rational behavior, I expect to be able to identify behavior sequences which are currently inexplicable by the system. Provided, that the required mental structure remains unchanged, I should be able to modify he considered domain of the mental states to optimize the system's predictions.

After this first step, I plan to explore a range of different assumptions, such as assuming that the system can correctly replicate the agent's perspective of the environment, required to infer necessary changes to the mental structure as well.

Finally, I plan to test the results from my thesis in a simple interaction game: The game will take place in an environment the system might not be familiar with, including different sources of uncertainty. A user and the system will then both control an agent in the environment and solve either cooperative or competitive tasks in which the system would be required to understand the other's behavior in order to be successful. An example would be that the human agent tries to reach a certain goal, but some ways may be blocked unknown to him. The system should ideally infer the agent's goal and provide useful information or help by removing the obstacle in time.

4 CONCLUSION

Summarizing, my thesis focuses on how artificial systems can be equipped with the capabilities to learn about the mental states of others well enough to explain their behavior in an efficient fashion in changing environments. My results will hopefully enable artificial systems to employ mentalizing more easily in a wide range of human-machine interaction scenarios, thereby qualitatively improving these interactions.

REFERENCES

[1] Chris Baker, Rebecca Saxe, and Joshua Tenenbaum. 2011. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the Cognitive Science Society*, Vol. 33.

[2] Chris L Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B Tenenbaum. 2017. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour* 1, 4 (2017), 0064.

[3] Chris L. Baker, Rebecca R. Saxe, and Joshua B. Tenenbaum. 2009. Action understanding as inverse planning. *Cognition* 113, 3 (2009), 329–349. https://doi.org/10.1016/j.cognition.2009.07.005

[4] Eugene Charniak and Robert P. Goldman. 1993. A Bayesian model of plan recognition. *Artificial Intelligence* 64, 1 (1993), 53–79.

[5] Harmen de Weerd, Rineke Verbrugge, and Bart Verheij. 2015. Higher-order theory of mind in the tacit communication game. *Biologically Inspired Cognitive Architectures* 11 (2015), 10–21.

[6] Andreea O. Diaconescu, Christoph Mathys, Lilian A E Weber, Jean Daunizeau, Lars Kasper, Ekaterina I. Lomakina, Ernst Fehr, and Klaas E. Stephan. 2014. Inferring on the Intentions of Others by Hierarchical Bayesian Learning. *PLoS Computational Biology* 10, 9 (2014). https://doi.org/10.1371/journal.pcbi.1003810

[7] Lawrence A Hirschfeld and Susan A Gelman. 1994. *Mapping the mind: Domain specificity in cognition and culture.* Cambridge University Press.

[8] Alan Jern, Christopher Lucas, and Charles Kemp. 2017. People learn other people's preferences through inverse decision-making. (2017).

[9] Daniel Kahneman. 2003. A perspective on judgment and choice: mapping bounded rationality. *American psychologist* 58, 9 (2003), 697.

[10] Ross A Knepper, Christoforos I Mavrogiannis, Julia Proft, and Claire Liang. 2017. Implicit communication in a joint action. In *Proceedings of the 2017 acm/ieee international conference on human-robot interaction.* ACM, 283–292.

[11] Shari Liu, Tomer D. Ullman, Joshua B. Tenenbaum, and Elizabeth S. Spelke. 2017. SM: Ten-month-old infants infer the value of goals from the costs of actions. *Science* 358, 6366 (2017), 1038–1041. https://doi.org/10.1126/science.aag2132

[12] Wulf-Uwe Meyer, Rainer Reisenzein, and Achim Schützwohl. 1997. Toward a process analysis of emotions: The case of surprise. *Motivation and Emotion* 21, 3 (1997), 251–274.

[13] Jan Pöppel and Stefan Kopp. 2018. Satisficing Models of Bayesian Theory of Mind for Explaining Behavior of Differently Uncertain Agents. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems.*

[14] David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences* 1, 4 (1978), 515–526.

[15] David V Pynadath and Stacy Marsella. 2007. Minimal mental models. In *AAAI.* 1038–1044.

[16] D. Ullman, Tomer, Chris L. Baker, Owen Macindoe, Noah D. Goodman, and Joshua B. Tenenbaum. 2009. Help or Hinder: Bayesian Models of Social Goal Inference. *Nips* (2009), 1–9. https://papers.nips.cc/paper/3747-help-or-hinder-bayesian-models-of-social-goal-inference.pdf

[17] Joey Velez-Ginorio, Max Siegel, Joshua B. Tenenbaum, and Julian Jara-Ettinger. 2017. Interpreting actions by attributing compositional desires. (2017).

[18] Henry M Wellman and David Liu. 2004. Scaling of theory-of-mind tasks. *Child development* 75, 2 (2004), 523–541.