Exploiting Attention to Reveal Shortcomings in Memory Models

Kaylee Burns
UC Berkeley
kayleeburns@berkeley.edu

Aida Nematzadeh
DeepMind
nematzadeh@google.com

Erin Grant
UC Berkeley
eringrant@berkeley.edu

Alison Gopnik
UC Berkeley
gopnik@berkeley.edu

Thomas L. Griffiths
Princeton University
tomg@princeton.edu

Abstract

The decision making processes of deep networks are difficult to understand and while their accuracy often improves with increased architectural complexity, so too does their opacity. Practical use of machine learning models, especially for question and answering applications, demands a system that is interpretable. We analyze the attention of a memory network model to reconcile contradictory performance on a challenging question-answering dataset that is inspired by theory-of-mind experiments. We equate success on questions to task classification, which explains not only test-time failures but also how well the model generalizes to new training conditions.

1 Reasoning about Beliefs

Possessing a capacity similar to human reasoning has been argued to be necessary for the success of artificial intelligence systems (*e.g.*, Levesque et al., 2011). One well-studied domain that requires reasoning is question answering, where simply memorizing and looking up information is often not enough to correctly answer a question.

Recent research has focused on developing neural models that succeed in such scenarios (Sukhbaatar et al., 2015; Henaff et al., 2017). As a benchmark to evaluate these models, Weston et al. (2016) released a dataset – Facebook bAbi – that provides a set of toy tasks, each examining a specific type of reasoning. However, the bAbi tasks are already too simple for the current models, which fail at only one or two (out of 20) tasks (Rae et al., 2016; Santoro et al., 2017).

Considering humans' reasoning abilities can provide inspiration for more complex tasks. People reason not just about their own observations and beliefs but also about others' mental states (such as beliefs and intentions). The capacity to recognize that others can have mental states different than one's own – *theory of mind* – marks an important milestone in the development of children and has been extensively studied by psychologists (for a review, see Flavell, 2004). Recently, Nematzadeh et al. (2018) released a dataset inspired by the theory-of-mind experiments from Baron-Cohen et al. (1985). The dataset is based on three tasks designed to capture increasingly complex theory-of-mind reasoning: *true-*, *false-*, and *second-order false-belief* tasks. Examples of each task type are given in Figure 1. In the true-belief task, Sally observes the world and as a result she has a first-order *true-belief* about the location of the milk – her belief matches reality. In the false-belief task, Sally's first-order belief differs from reality (*i.e.*, she has a *false-belief*) because she was absent when the state of the world changed. In the second-order false-belief task, Sally observes the new location of the milk; thus, she has a *true-belief* about the milk's location. However, Anne's belief about Sally's mental state does not match reality because Anne does not know that Sally has observed the change in the environment. As a result, Anne has a *false belief* about Sally's beliefs.

The dataset from Nematzadeh et al. (2018) contains 4 question types: 2 related to world state and 2 related to beliefs (Table 1). These questions enable us to test whether a model can reason about first-order and second-order beliefs and know the initial and current location of an object; thus, we can distinguish between when a model answers a question by chance and when it actually understands the entire state of the world. Table 2 gives the answers for the 12 combinations of task type and question. Our analysis will focus on the two belief questions proposed.

We use these tasks to generate a training set with 10 000 examples with each of the 12 combinations of task and question types, randomly

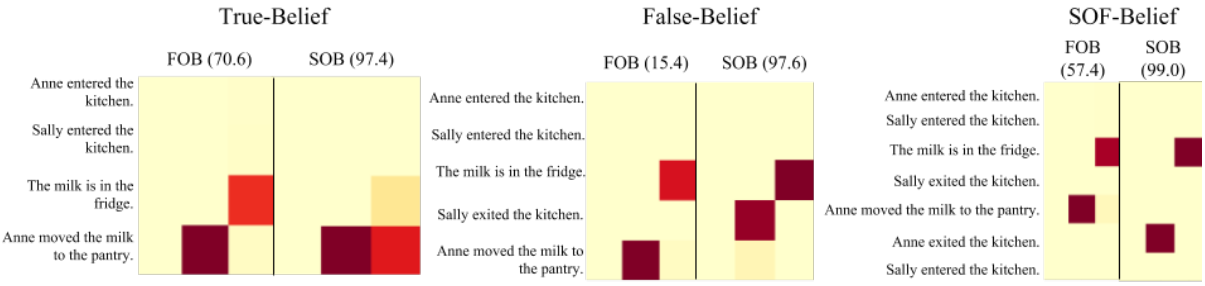

Figure 1: The attention of MemN2N in response to first-order (all left) and second-order (all right) belief questions. To correctly answer the second order belief question, the model needs to identify the true belief task from other tasks (see Table 2). To this end, the model can use the presence of "exit" to classify true-belief from non-true-belief tasks. There is no analogous identifier for the first-order question, where the model fails.

Memory	Where was the milk at the beginning?
Reality	Where is the milk really?
First-order	Where will Sally look for the milk?
Second-order	Where does Sally think that Anne searches for the milk?

Table 1: Examples of the four question types.

	True Belief	**False Belief**	**Second-Order FB**
Memory	first	first	first
Reality	second	second	second
First-order	second	first	second
Second-order	second	first	first

Table 2: The correct answer to each question. Here, "first" and "second" are the the initial and actual locations of the object of interest.

grouped into sets of 5 to form stories. Each story in the test set contains 4 tasks, but there is only one question present at the end. Because questions that come closer to the beginning of a story have fewer distractors (*i.e.*, potential answer words) that may confound a model, they are easier to answer.

2 Experiments

We train MemN2N (Sukhbaatar et al., 2015) jointly over all task types without noise, but evaluate success on a test set with noise sentences generated randomly at different positions (*i.e. ToM* (noised)). We first examine how the model performs across a range of parameter and initialization values. Because MemN2N models are very sensitive to the network initialization, for each set of parameters, the best result out of 10 runs is used for each configuration of hyperparameters. To understand why failures occur, we plot the average attention over all instances of each task-question combination. Figure 1 shows the average attention of the best performing 3-hop model on the first-order (left) and second-order (right) belief tasks. Only the attention over memory slots with relevant story sentences is displayed.

Surprisingly, the model is successful on the "harder" second-order belief question but not on the first-order one. Indeed, the pattern of attention across hops in response to the second-order belief question is more varied across task conditions and attends to sentences that provide information about agents' transition in the world (*i.e.*, "Sally exited the kitchen"). On the other hand, the left hand side of the figure shows that, in response to the first-order belief question, the attention is not sensitive to the task type (*i.e.*, true-, false- or second-order- belief).

Considering each belief question as a task classification, as shown in Table 2, can explain this result. The answer to the first-order question is different for false-belief and second-order false-belief tasks while it is the same for the second-order question. Given the similarity of these 2 tasks (*e.g.*, Sally moves between rooms in both tasks, both contain the word "exited"), the "classification" problem is much easier when the two questions have the same answer. To answer the first-order question correctly – where the answers are different for the false-belief and second-order false-belief tasks – the model needs to learn to distinguish between between these very similar tasks.

To further test this hypothesis, we created an inaccurate version of the *ToM* dataset where the answer to the false belief question was modified to be the second location of the object as opposed to the first. With the difficulty of classifying false-belief from second-order false belief tasks removed, the models were able to successfully answer all of the first order belief questions.

References

Simon Baron-Cohen, Alan M Leslie, and Uta Frith. 1985. Does the autistic child have a theory of mind? *Cognition*, 21(1):37–46.

John H Flavell. 2004. Theory-of-mind development: Retrospect and prospect. *Merrill-Palmer Quarterly*, 50(3):274–290.

Mikael Henaff, Jason Weston, Arthur Szlam, Antoine Bordes, and Yann LeCun. 2017. Tracking the world state with recurrent entity networks. In *International Conference on Learning Representations*.

Hector J Levesque, Ernest Davis, and Leora Morgenstern. 2011. The winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, volume 46, page 47.

Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Thomas L. Griffiths. 2018. Evaluating theory of mind in question answering. In *Conference on Empirical Methods in Natural Language Processing*, volume abs/1604.06045.

Jack W. Rae, Jonathan J. Hunt, Tim Harley, Ivo Danihelka, Andrew W. Senior, Greg Wayne, Alex Graves, and Timothy P. Lillicrap. 2016. Scaling memory-augmented neural networks with sparse reads and writes. In *Proceedings of 30th Conference on Neural Information Processing Systems*.

Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy P. Lillicrap. 2017. A simple neural network module for relational reasoning. In *International Conference on Learning Representations*.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *Proceedings of 29th Conference on Neural Information Processing Systems*, pages 2440–2448.

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. 2016. Towards AI-complete question answering: A set of prerequisite toy tasks. In *International Conference on Learning Representations*.