FAITHNET: A GENERATIVE FRAMEWORK IN HUMAN MENTALIZING

Chengkai Guo*
ggsonic@gmail.com

August 29, 2022

ABSTRACT

In this paper, we first review some of the innovations in modeling mentalizing.Broadly, this involves building models of computing World Model and Theory of Mind(ToM). A simple framework, FaithNet, is then presented with concepts like persistence, continuity, cooperation and preference represented as faith rules.FaithNet defines a generative model that can sample faith rules. Our FaithNet utilize a general-purpose conditioning mechanism based on cross-attention, offering computations that best explain observed real-world events under a Bayesian criterion. Our code will be released at `https://github.com/ggsonic/GIN`

1 Introduction

Recent progress in artificial intelligence (AI) has growing interest in building AGI systems that learn and think like people. Many advances have come from using deep learning neural networks trained end-to-end in tasks such as objectre cognition, video games, and board games, achieving even superhuman performance in some respects. Despite their biological inspiration and performance achievements, contemporary AI (and deep learning models in particular) solve problems in different ways than people do.

There are two general approaches to computational intelligence,the statistical way and the mentalizing way . The statistical pattern recognition approach treats prediction as primary, usually in the context of a specific classification, regression, or control task. In this view, learning is about discovering features that have high value states in common – a shared label in a classification setting or a shared value in a reinforcement learning setting – across a large, diverse set of training data. The other approach treats models of the world (and the mind) as primary, where learning is the process of model-building. Cognition is about using these models to understand the world, to explain what we see, to imagine what could have happened that didn't, or what could be true that isn't, and then planning actions to make it so. The difference between pattern recognition and model-building, between prediction and explanation, between statistics and mentalizing, is central to our view of human intelligence. Just as scientists seek to explain nature, not simply predict it, we see human thought as fundamentally a model-building activity. We elaborate this key point with numerous examples below. We also discuss how pattern recognition, even if it is not the core of intelligence, can nonetheless support model-building, through "model-free" algorithms that learn through experience how to make essential inferences more computationally efficient.

In cognitive science we argue that truly human-like learning and thinking machines should (a) build causal models of the world that support explanation and understanding, rather than merely solving "prediction" problems; (b) ground learning in intuitive theories of physics and psychology, to support and enrich the knowledge that is learned; and (c) harness compositionality and learning-to-learn to rapidly acquire and generalize knowledge to new tasks and situations. We also suggest promising routes towards these goals that can combine the strengths of recent neural network advances with more structured cognitive models.

In this paper,a simple framework, FaithNet, is presented with concepts like persistence, continuity, cooperation and preference represented as faith rules. Our FaithNet Framework defines a generative model that can sample faith rules

*This work was performed while the author was an independent researcher

and utilize a general-purpose conditioning mechanism based on cross-attention, offering computations that best explain observed real-world events under a Bayesian criterion.

2 Developmental Psychology

Mental state inference (or 'mentalizing') in adults probably draws on a diverse set of representations and processes, but our focus is on a capacity that appears in some form in infancy [1] and persists as a richer theory of mind develops through the first years of life [2]. What we call core mentalizing is grounded in perception, action and the physical world: it is based on observing and predicting the behaviour of agents reaching for, moving toward or manipulating objects in their immediate spatial environment, forming beliefs based on what they can see in their line of sight, and interacting with other nearby agents who have analogous beliefs, desires and percepts. In contrast to more explicit, language-based theory-of-mind tasks, which are only passed by older children, these core abilities can be formalized using the math of perception from sparse noisy data and action planning in simple motor systems. Hence, core mentalizing is an aspect of social cognition that is particularly likely to be readily explained in terms of rational computational principles that make precise quantitative predictions.

2.1 Intuitive physics

Young children have rich knowledge of intuitive physics. Whether learned or innate, important physical concepts are present at earlier ages. At the age of 2 months and possibly earlier, human infants expect inanimate objects to follow principles of persistence, continuity, cohesion and solidity. Young infants believe objects should move along smooth paths, not wink in and out of existence, not inter-penetrate and not act at a distance [3]. These expectations guide object segmentation in early infancy, emerging before appearance-based cues such as color, texture, and perceptual goodness. At around 6 months, infants have already developed different expectations for rigid bodies, soft bodies and liquids [4]. Liquids, for example, are expected to go through barriers, while solid objects cannot [5]. By their first birthday, infants have gone through several transitions of comprehending basic physical concepts such as inertia, support, containment and collisions [6].

2.2 Intuitive psychology

The development of theory of mind can help us to understand the learning phases of ToM skills and strategies. It is considered that by the age of 5 children have developed many aspects of ToM. From around 6 months of age [7] human infants begin to distinguish between the motion of inanimate and animate objects. At 12 months of age joint attention is developed, where the infant has the cognitive capacity to represent its own perception, that of an agent (e.g., mother) and that of an object. By 14–18 months, through gaze direction, the infant begins to understand the mental states of desire, intention and the causal relation between emotions and goals [8]. Liszkowski et al. [9] showed that children as young as 12–18 months were able to infer an adult's behavior and aid them. In this particular experiment, infants watched an adult write with a marker on a piece of paper. The marker would drop off the table, not seen by the adult. When the adult began randomly searching for the marker, the infant would either point to or retrieve the marker, ignoring any other distractors. Between 18 and 24 months toddlers begin to distinguish between real and pretend events and often start to engage in pretend play around this age. Around the age of 3–4 children begin to understand the differences between their own and others' beliefs and knowledge, thereby beginning to comprehend false beliefs, but this ability does not become fully stable until age 5–6. Understanding metaphors, irony and sarcasm only establishes around age 6–7.

3 Computational models

Computational models provide a way to formally operationalize the mental processes that are hypothesized to underlie a specific cognitive operation. This allows us to simulate behavior in various tasks and assess if the model can accurately account for how people behave in these environments.

3.1 Reinforcement Learning

RL provides a framework to not only learn about the moral character of another person, but also about the world via another person's experience. While observing an agent interact with their environment, prediction errors about the outcome can be computed for the agent rather than the observer.

An observer can also learn to imitate which actions to take, even when the outcomes or rewards for an observed agent's actions are not directly observable. Rather than learning the reward contingencies of the environment, imitation learning involves learning to take a particular action based on the extent to which another agent was observed to take that action in the past. Here, the value of a given action is computed through positive reinforcement if the action was performed by an observed agent, while unchosen actions are negatively reinforced.

Unlike standard RL, which attempts to learn the optimal actions given a reward function, inverse reinforcement learning, attempts to recover the learned reward function for which an observed agent's actions would be optimal. This type of algorithm is particularly well suited to model hidden beliefs, goals, and desires from observing others' actions.

3.2 Bayesian Theory of Mind

A criticism of using RL models as a model of rational behavior, however, is that agents in the real world rarely possess full knowledge of their environment. Partially observable Markov decision processes (POMDPs) attempt to model the causal relationship between an agent's beliefs and their actions given their uncertainty about the state of the world.

POMDPs capture three central causal principles of core mentalizing : a rational agent (I) forms percepts that are a rational function of the world state, their own state and the nature of their perceptual apparatus — for a visually guided agent, anything in their line of sight should register in their world model (perception); (II) forms beliefs that are rational inferences based on the combination of their percepts and their prior knowledge (inference); and (III) plans rational sequences of actions — actions that, given their beliefs, can be expected to achieve their desires efficiently and reliably (planning). BToM [10] integrates the POMDP generative model with a hypothesis space of candidate mental states, and a prior over those hypotheses, to make Bayesian inferences of beliefs, desires and percepts, given an agent's behaviour in a situational context.

In the single action case, given the prior $Pr(B_0, D, P, S)$ over the agent's initial beliefs B_0, desires D and the situation S, the likelihoods defined by principles (I–III) above, and conditioning on observations A of how the agent then acts in that situation, the BToM observer can infer the posterior probability $Pr(B, D, P, S \mid A)$ of mental states (belief states $B = B_0, B_1$, desires D, and percepts P), and the situation S given actions A using Bayes' rule:

$$Pr(B, D, P, S \mid A) \propto \quad Pr\left(A \mid B_1, D\right) \times Pr\left(B_1 \mid P, B_0\right) \times Pr(P \mid S) \times Pr\left(B_0, D, S\right) \tag{1}$$

The BToM model formalizes mentalizing as Bayesian inference over a generative model of a rational agent. BToM defines the core representation of rational agency using partially observable Markov decision processes (POMDPs).

4 FaithNet

This paper introduces the FaithNet framework, capable of learning a large class of mentalizing rules from daily experience and generalizing to wild tasks through generative models.

4.1 Method

In our FaithNet framework,Mentalizing are represented as simple faith rules with conditional generative models.Whether inherited or learned from experience,faith rules are consistent with essential human-level concepts like persistence, continuity, cooperation and preference,etc. And a general-purpose conditioning mechanism based on cross-attention is used to build the generative model. Our FaithNet framework brings together three key ideas—compositionality, causality, and learning to learn—that have been separately influential in cognitive science and machine learning over the past several decades. As conditional generative models, rich mentalizing can be built "compositionally" from simpler faith rules. Their probabilistic semantics handle noise and support creative generalizations naturally captures the abstract "causal" structure of the real-world processes. Learning proceeds by constructing faith rules that best explain the observations under a Bayesian criterion, and the model "learns to learn" by developing hierarchical priors. In short, FaithNet can reasoning new real-world events by reusing the faith rules of existing ones, capturing the causal and compositional properties of real-world generative processes operating on multiple scales.

The FaithNet learns simple faith rules to represent real-world events, building them compositionally from rules and spatial-temporal relations. FaithNet defines a generative model that can sample faith rules . The joint distribution on events ϵ, a set of M rules of that type $\kappa^{(1)}$,..., $\kappa^{(M)}$, and the corresponding observations $o^{(1)}$,...,$o^{(M)}$ factors as:

$$P\left(\epsilon, \kappa^{(1)}, \ldots, \kappa^{(M)}, o^{(1)}, \ldots, o^{(M)}\right) = P(\epsilon) \prod_{m=1}^{M} P\left(o^{(m)} \mid \kappa^{(m)}\right) P\left(\kappa^{(m)} \mid \epsilon\right) \tag{2}$$

4.2 Network Architecture

To lower the computational demands of training FaithNet models, we observe that although many work have been done on exploring deep learning generative models (GAN,VAE,Diffusion models), they still require costly function evaluations in pixel space, which causes huge demands in computation time and energy resources. We adress this drawback with our proposed FaithNet, which work on a compressed latent space of lower dimensionality. Our FaithNet utilize a general-purpose conditioning mechanism based on cross-attention, enabling multi-modal training. Faith rules learned as latent codes are mapped into the UNet via (multi-head) crossattention with contextual types of conditionings,which are then mapped to the intermediate layers of the UNet via a cross-attention layer implementing. Such an approach offers several advantages: (i) By leaving the high-dimensional image space, we obtain faithnet which are computationally much more efficient because sampling is performed on a low-dimensional space. (ii) We exploit the inductive bias of faithnet inherited from their UNet architecture, which makes them particularly effective for data with spatial structure and therefore alleviates the need for aggressive, quality-reducing compression levels as required by previous approaches . (iii) Finally, we obtain general-purpose compression models whose latent space can be used to train multiple generative models.

Once a low-level faith rule has been trained to model a large variety of real-world events, it can then be reused to solve new events. The pre-trained low-level faith rule enable our faithnet to produce high-level behaviors using only simple task-reward functions.

5 Discussion

We have briefly reviewed the innovations within computing mentalizing domains to provide a general framework to model the essential dynamics of World Model, modules for how the agent might represent the mental states of other agents, such as their beliefs, goals, desires, intentions, and feelings, and modules for how to integrate internal goals and mentalizing computations to produce optimal policies to navigate the environment. The strengths of this computational approach is that the framework and mathematical operationalization of these constructs facilitates collaborations across different laboratories and also scientific disciplines.

In our FaithNet framework, under the constraints of minimizing information entropy, the number of faith rules is very small and highly abstract. The same faith can generalize to the behavior of different entities, living entities, abstract concepts, social groups, etc. under different conditions. And it can be continuously compressed and optimized with the increase of individual experience, making the calculation faster and the generalization ability more advanced and abstract. It is even solidified into a neuron-level composition and passed on to future generations to achieve knowledge or ability-level transmission.

References

[1] A. L. Woodward. Infants selectively encode the goal object of an actor's reach. *Cognition 69, 1–34*, 1998.

[2] H. Wimmer and J. Perner. Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition 13, 103–128*, 1983.

[3] E. S. Spelke. Principles of object perception. *Cognitive Science, 14 (1), 29–56.*, 1990.

[4] L. J. Rips and S. J. Hespos. Divisions of the physical world: Concepts of objects and substances. *Psychological Bulletin, 141 , 786–811.*, 2015.

[5] S. J. Hespos, A. L. Ferry, and L. J. Rips. Five-month-old infants have different expectations for solids and liquids. *Psychological Science, 20 (5), 603–611.*, 2009.

[6] R. Baillargeon. Infants' physical world. *Current Directions in Psychological Science, 13 , 89–94.*, 2004.

[7] S. Baron-Cohen. Mindblindness: An essay on autism and theory of mind. *Cambridge, MA: MIT Press*, 1995.

[8] R. Saxe, S. Carey, and N. Kanwisher. Understanding other minds: linking developmental psychology and functional neuroimaging. *Annu. Rev. Psychol. 55, 87–124.*, 2004.

[9] U. Liszkowski, M. Carpenter, T. Striano, and M. Tomasello. 12-and 18-month-olds point to provide information for others. *J. Cogn. Dev. 7, 173–187.*, 2006.

[10] 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, 1(4), s41562–017 – 0064.*, 2017.