Workshop proposal: Deep Learning in Computational Cognitive Science

Ilker Yildirim (ilkery@mit.edu) and Joshua B. Tenenbaum (jbt@mit.edu)
Department of Brain & Cognitive Sciences, MIT
Cambridge, MA 02130 USA

Keywords: Computational models of cognition; deep learning; Bayesian models; cognitive neuroscience; computational neuroscience; computational psycholinguistics.

Overview and significance

A new generation of deep neural network architectures has driven rapid advances in AI over the last ten years. These architectures include convolutional neural networks (CNNs), recurrent neural networks (RNNs), and many variants and extensions. Computational cognitive scientists and neuroscientists have now begun to explore these techniques, and how they might combine with other computational tools such as Bayesian models, symbolic grammars and rule-systems, probabilistic programs, and reinforcement learning. The goal of this workshop is to bring together some of the leading researchers working at this interface, for short talks and an integrative discussion of open questions and promising directions.

Talks will cover many areas of cognition including perception, problem-solving and planning, decision-making, language and social cognition. The focus will be on models of human behavior, but the potential bridge to neural studies in humans (via fMRI) and animals (via physiology) will also be explored. Most talks will assume only a basic familiarity with neural networks, and so should be accessible to all CogSci attendees. *We hope to be scheduled for an afternoon slot, and have coordinated our plans with the DeepMind's Deep Learning tutorial proposed for the morning which could serve as an introduction to more advanced methods that several talks will build on.*

Workshop structure. We plan a half-day workshop comprising seven talks, each 20-25 minutes, followed by a 30-minute panel discussion with all speakers on open questions. We will also encourage student participants to present posters on relevant work during the coffee break.

Organizers and Presenters

Ilker Yildirim (Organizer) is a research scientist at MIT. His research spans visual and multisensory perception, computational neuroscience, and artificial intelligence

Joshua Tenenbaum (Organizer) is Professor of Computational Cognitive Science at MIT. He studies learning, perception, common-sense reasoning, and has been active in both cognitive science and artificial intelligence.

Matt Botvinick is DeepMind's Director of Neuroscience Research, and was formerly Professor of Psychology and Neuroscience at Princeton. He is a leader in computational cognitive neuroscience and reinforcement learning.

Noah Goodman is a Associate Professor of Psychology, Linguistics and Computer Science at Stanford University. His research centers on computational modeling of higher-level cognition and probabilistic programming languages.

Thomas Griffiths is a Professor of Psychology and Cognitive Science at UC Berkeley. His group develops computational models of higher-level cognition, drawing on probabilistic, neural network, and evolutionary paradigms.

Jessica Hamrick is a PhD candidate at UC Berkeley and former intern at DeepMind. Her research focuses on mental simulation, planning and metacognition.

Tal Linzen is a postdoctoral researcher at ENS and will be Assistant Professor in the Department of Cognitive Science at Johns Hopkins starting in the fall. His research interests involve computational modeling and psycholinguistics.

Daniel Yamins is an Assistant Professor of Computational Neuroscience in the Department of Psychology at Stanford. His research lies at the intersection of neuroscience, artificial intelligence, and psychology.

Presentations

Prefrontal cortex as a meta-reinforcement learning system (Botvinick). Two decades of neuroscience research on reward-based learning has converged on a canonical model, under which the neurotransmitter dopamine 'stamps in' associations between situations, actions and rewards by modulating the strength of synaptic connections between neurons. However, a growing number of recent findings have placed this standard model under strain. This talk draws on recent advances in AI to introduce a new theory of reward-based learning. Here, dopamine trains another part of the brain, the prefrontal cortex, to operate as its own free-standing learning system. This perspective accommodates the findings that motivated the standard model, but also deals gracefully with a wider range of observations, laying a fresh foundation for next-generation research.

Less-supervised loss functions for training models of the visual system (Yamins). Recent advances in computer vision and AI have made it possible to build deep neural networks that mimic aspects of computation in the primate and human visual system. The core idea behind these results is task-driven modeling, e.g. optimize a neural network for a complex ecologically relevant behavior, and then compare the learned model to neural responses in the brain areas thought to underlie that behavior. While this approach has produced powerful predictive models of neural representations in adult animals, its main successes so far have unfortunately relied on using heavy semantic

supervision, using large labeled datasets – data streams to which real animals do not have access. I will discuss recent work my lab has been doing to move beyond this limitation, developing less heavily supervised approaches to train deep neural network models of vision that may be more plausible models of real neural learning. These ideas rely on loss functions defined in interactive worlds that attempt to better capture the true complexities of the environment in which early juvenile development takes place.

Modeling "analysis by synthesis" in perception by combining generative models and deep inverse networks (Yildirim and Tenenbaum). In its most general form, perception can be defined as the solution to an inverse problem: identifying the world scene that gave rise to the observed retinal (or auditory, or haptic) data. It remains a mystery how the brain constructs such rich representations of the geometry of objects in a scene and their physical properties extremely quickly, in at most a few hundred milliseconds. Traditional models of how the brain solves such inference problems use iterative methods that are hard to map onto neural circuits and much too slow to explain online perception. Here we present a new approach that combines a deep neural network with a probabilistic generative model. This approach is as fast as pure feed-forward models, but generates a rich description of 3D object shapes and physical properties. Applied to faces, our model explains both the tuning properties of cells in the macaque face patch system and human behavior in recognizing familiar and unfamiliar individuals.

Deep networks for amortized inference in structured probabilistic models (Goodman). I will discuss deep amortized inference for probabilistic programming languages (PPLs). This is an approach that uses a PPL to describe complex probabilistic conceptual knowledge, and captures knowledge about how to use these concepts for inference via a deep 'inference network'. I will discuss several different objectives and training methods, including variational inference and 'dream learning' (a modern variant of wake/sleep). After describing the technical setup and showing the results for a few model-learning tasks, I will speculate about the relation of deep inference networks to human procedural knowledge.

Leveraging deep learning to study representations underlying human cognition (Griffiths). Recent neural network models have resulted in significant progress in computer vision, speech recognition, and natural language processing, by learning representations of the statistical structure of complex visual, auditory, and linguistic stimuli. Understanding how the resulting models work — and how well they correspond with human perception — is an interesting scientific challenge. However, the representations that these models discover, when treated just as representations of complex stimuli, also offer the opportunity to extend the scope of psychological research. Psychologists studying problems such as categorization or memory have tended to focus on very simple stimuli that can be carefully controlled and parameterized. This makes it possible to formulate precise theories, but at the cost of potentially losing sight of the original phenomena: Do the same models that predict how people categorize sinusoidal gratings explain how they differentiate cats and dogs? I will talk about recent work that tries to leverage representations produced by neural networks — for both images and language — to study human cognition, highlighting the promise of this approach as well as some of the challenges.

Metacontrol for Adaptive Imagination Based Optimization (Hamrick). Many machine learning systems are built to solve the hardest examples of a particular task, which often makes them large and expensive to run---especially with respect to the easier examples, which might require much less computation. For an agent with a limited computational budget, this "one-size-fits-all" approach may result in the agent wasting valuable computation on easy examples, while not spending enough on hard examples. Rather than learning a single, fixed policy for solving all instances of a task, we introduce a "metacontroller" inspired by human cognition which learns to optimize a sequence of imagined internal simulations over predictive models of the world (called "experts") in order to construct a more informed, and more economical, solution. Our approach learns to adapt the amount of computation it performs to the difficulty of the task, as well as which experts to consult by factoring in both their reliability and individual computational resource costs. The metacontroller achieves a lower overall cost (task loss plus computational cost) than more traditional fixed policy approaches, demonstrating that our approach is a powerful framework for using rich forward models for efficient model-based reinforcement learning.

Understanding neural models of human language (Linzen). Large-scale artificial neural networks have shown great promise in natural language processing, reportedly reaching human-level performance in some tasks. Yet our understanding of the capabilities of these methods is typically limited to general statistics averaged across a random sample of texts. Such coarse-grained evaluation metrics stand in marked contrast to the rich array of highly specific patterns identified by linguists and cognitive scientists. I will argue that this detailed characterization of human-level knowledge of language provides a yardstick for the desired behavior of an artificial intelligence system. Applying it to neural networks can help us understand the strengths and weaknesses of existing architectures and analyze models that are otherwise difficult to interpret. Finally, neural networks that combine powerful statistical learning with different degrees of representational assumptions can serve as useful baselines for psychological modeling. I'll discuss case studies illustrating this approach in both syntax and semantics.