Neuroscience-Inspired Artificial Intelligence

Reminders

- Marr's levels of understanding complex biological systems (higher to lower):
  1. Computation level - what the goal of the system is,
  2. Algorithm level - how is that goal achieved (through which processes),
  3. Implementation level - how are these algorithms implemented on a low level

Overview

- Connection between neuroscience and artificial intelligence was always evident, and more so in the beginning of AI development
- Taking inspiritaion from biological intelligence has two important benefits:
  - new algorithms and architectures can be discovered from inspiration from neuroscience,
  - AI algorithms can be validated if found in biological systems
- This paper focuses on top two Marr's levels of understanding of the brain - looking at the most prominent fields of AI research Deep and Reinforcement Learning

Key ingredients

- *Exploring the potential benefits of NS to AI*

Past

- Deep Learning
  - Origins lie in NS, as the name "neural networks" suggests
  - Milestone in discovering backpropagation for MLPs in 1985
  - Parallel Distributed Processing movement (1985) - moved from the idea of brain working sequentially, to the idea of stochastic, highly parallelized information processing (backed by research in NS)
  - Strong examples of inspirations from NS:
    * Mammalian visual cortex reveals how visual input is filtered and pooled, much like in CNNs
    * Both convergent and divergent information flow in successive processing layers, replicated in current NNs
    * Dropout used to model stochasticity
- Reinforcement Learning
  - Initially inspired by research into animal learning; particularly temporal-difference (TD) learning
  - TD learning explains the phenomena of second-order conditioning, where the learning is associated to another conditioned stimulus instead with the unconditioned one (e.g. Q to Q', instead of directly to expected return)

Present

- Attention
  - Comes from an important insight that the **biological brains are modular**, with distinct interacting subsystems
  - Change from looking at the whole image, to selecting specific parts to look at - the same as in the biological brain, the attention shifts among locations and objects to in order to save resources
  - Can also be used internally, towards memory, e.g. by selecting which information to read from the internal memory of the network - used for machine translation
  - Generative models use attention to iteratively generate outputs (e.g. DRAW for image generation)
- Episodic memory
  - Prominent theory suggests that the learning is done by complementary learning systems in hippocampus and neocortex:

    * hippocampus encodes new information (one-shot learning), but is not able to generalize and is non-parametric
    * neocortex slowly learns (consolidates information during resting), has generalizing capabilities, and is parametric
  - Replay buffer in RL helps stabilize learning and avoid catastrophic forgetting that happens with change in input distributions
    * replay in hippocampus seems to favor events that lead to higher level of reinforcement (prioritized replay)
- Working memory
  - Thought to be instantiated in the prefrontal cortex
  - RNNs/LSTMs draws ideas in the early work in NS, where sequence control and memory storage are closely intertwined - new theories suggest they are separated
  - Differential Neural Computer addresses new theories better, with separated sequential control and memery, and read/write capabilities
- Continual learning
  - important to be able to learn new things while retaining previous knowledge
  - in humans forgetting is shown to be prevented by specialized mechanisms that decrease plasticity in previously learned tasks
  - inspired by this Elastic Weight Consolidation networks are implemented that identify important weights and slow for previous tasks and slow down their updates

Future

- Intuitive understanding of the physical world
- Efficient learning
  - e.g. learning to learn, or leveraging prior knowledge
- Transfer learning
- Imagination and planning
  - imagining mental models and planning
  - examples in RL - Dyna and MCTS
  - NS suggests that hippocampus supports planning by creating an internal model of the world
- Virtual brain analytics
  - analyzing the brain by applying tools from AI and NS to increase understanding

AI to NS

- ML contributions to different fields, e.g. MRI for NS
- RL in understanding TD learning
- CNN insights towards understanding visual areas
- LSTM for insights that motivated development of working memory models
- Insights in understanding how the memory works
- Meta reinforcement learning insights in different learning speeds in humans
- Insights in how backpropagation works in humans

Comments

- Really like the points made which causally connect research in AI to RL
- A lot of terms in NS that I can benefit to
- Generally influenced me to want to get a better understanding of the NS approaches and dive deeper into the field
- Also a lot of good references, especially for older papers
- Good to know the insights for the future from the NS side