Research Proposal

My long-term research goal is to build an agent who can learn, create, and communicate like humans. By integrating perception, action, and language under a unified framework, embodied agents can lead us toward this ultimate goal of computer science. With an interdisciplinary background in computer science and neuroscience, I can uniquely contribute to this area.

Transformers - An Agent That Is Suitable for Any "Body"

The agent needs a set of outputs to interact with this world. The agent must learn to control a physical or a virtual body. Infants learn to control their bodies by trial and error. They dance randomly with their arms and legs to interact with the physical world. By searching in the state space, they learn what kind of action is possible and what's the outcome of this action. Motor control in humans can be described by a sequence of muscle commands. Our reinforcement learning network can also control a body by sequential decision-making with feedback from the inner proprioception and the outer world.

Before everything starts, the agent should perceive the world and its body states. Using unsupervised semantic segmentation, the agent can get the invariant representation of this 3D world. By exploration, the agent can learn a set of mapping from action sequences to state changes using RL. After that, the agent can learn new tasks by knowledge reusing and compositing. In addition, tools can be viewed as an extension of our bodies.

The Blind Men and the Elephant - Perception & Multimodality

We have multimodal perceptions to capture the information of this world, like vision, auditory, and tactile senses. This physical world can be measured in different forms. However, we can never touch the nature of any entities in this world. We only use sense organs to obtain partial information about our surroundings. Evolution helps us choose the most useful modality for survival, but not the truth. What's more, time and space do not objectively exist. They exist because we use these terms and dimensions to analyze this world. Thus, the agent should have enough information about this world by observation, but we should not constrain how it merges and use the information. The data from different modalities should be represented under a unified latent space. Current work about Transformers is an excellent example of unifying spatial and sequential information.

Another interesting question we can explore is why the agent should understand the dynamic 3D scene. As creatures, we develop our perception with purpose. So does the agent. Current research about active sensing and action-object binding can claim the purpose of 3D scene understanding. Action and Perception are inseparable. The agent can move around for a 3D object to observe it from different views. By locomotion, the agent builds the 3D inner model of entities in this world. Through physical interaction, the agent binding actions to objects.

Crystallized Intelligence - Long-term Memory

Memory consolidation happens during sleep. When we sleep, short-term memory can be

transformed and transported from short-term memory in the hippocampus to long-term memory in the cortex. Building blocks like LSTM and GRU mimicked such behavior successfully. Memory and information are restored through network connections. This means the connections patterns between neurons and the dynamic properties of neurons represent the information. From this perspective, the brain's architecture is in-memory computing instead of the traditional Von Neumann structure. Correspondingly, deep neural networks also use in-memory computing architecture. The connection weights inside the blocks represent the inductive bias of the task and the dataset. We can use incremental learning and life-long learning to update long-term memory.

The Imitation Game

Supervised learning is just like exam-oriented education. But newborn babies can learn by imitation. They may need to try thousands of combinations of actions to use a dinner fork. But with a demonstration, the baby can learn quickly. Imitation can speed up the exploration. The underlying mechanism of imitation learning is mirror cell, corollary discharge, and dopamine system. We can also achieve this in agents by action detection, object detection, 3D reconstruction, and inverse RL with egocentric or allocentric videos.

Mary in the Room - Embodied Experience

Our current AI is like Mary mentioned in the knowledge argument (also known as Mary's room or Mary the super-scientist). The agent is trained in the virtual world and learns everything it has never experienced. However, knowledge is grounded in our physical world as an abstraction of daily events. I believe that if agents never see red, they will never understand the true meaning of red. Many researchers are trying to address this problem, including grounded language, multimodal learning, text-to-image, and embodied robots. Intelligence is everything about compression, from concrete examples to abstract concepts. After that, the abstract concept can be manipulated to generate new ideas and instantiate into a particular case. This process works like an autoencoder or a generative network.

Our daily experience is the anchor of our consciousness. Robustness is a big challenge for current AI. The present agent does not have a "root" before developing intelligence. We may imagine we can fly in the air. But we can distinguish between reality and illusory. The world is under the rules of physics. People may not discover Newton's laws of motion, but they can still live happily and safely in this world. That's because we can build a latent inner model of this world. However, AI does not have embodied experience and common sense. Thus, they will make mistakes that people will not.

Language, Neuro-symbolic, and Time Travel

We can communicate because we have a shared understanding of this physical world and embodied experience. Our ancestors used images and hieroglyphics to record daily events. They generally evolved into abstract symbols. Language is deeply rooted in the concrete physical world. Symbols are meaningless by themselves. We give them meanings. By converting episodic memory into declarative memory, symbols and language allow us to recall past events without experiencing

“time travel” in our minds. Agents should also connect perception and action with language by using a higher-level network as a coordinator above multiplied asynchronous networks. Once agents can express themselves, understand human language, and learn by reading, we can cooperate with them efficiently and safely.

What’s the Meaning of the Agent’s Life - Target Function

We may work and live for our dream or our family. The agent also needs an optimization target. In the first stage, when the agent does not learn enough experience and can’t communicate like humans, its target function should be learning something new. The agent needs to build an inner model of this world. This requires the agent to explore this world and try everything new. The mutual information or cross-entropy between pre-existing knowledge and new-learned knowledge can measure novelty. After that, the agent needs to learn how to connect this inner model and perception network with symbols and languages. The agent can help humans to achieve their demands in this way.

Where Is the Boundary – Inherent Bias and Knowledge Acquisition

Another interesting question that I want to explore is how to draw a boundary between inherent structure and learned knowledge. For example, computers have predefined multi-level structures from the register, cache, and memory to disk. This structure is inherent. However, a computer can execute multiple tasks with this architecture. The program and data in the computer are acquired and encoded based on the architecture.

Regarding brain and deep learning, CNN and RNN have task biases for information with spatial and temporal modalities, respectively. This preexisting structure in our brain comes from evolution and is encoded in genes. Like the two-timing method in relaxation oscillation in a nonlinear dynamic system, biological dynamics should have various time scales. Evolution and gene are slow time scale factors. They are designed to fit the slow oscillation of the earth’s environment. Short-term memory and long-term memory are quick time scale factors. They respond to our highly dynamic and ever-changing environment.

An Agent Without Soul - Beyond Intelligence

What we have discussed above is all about intelligence. However, being a human is not about intelligence only. I know little about our emotions and humanity. We can build an agent capable of everything about logic using deep neural networks. But I am still trying to figure out what to do regarding philosophy. Our emotions are vague reactions. As for humanity, it might grow in our society. Just like intelligence may emerge from a complex system, humanity might also appear suddenly when a group of human-level agents interacts with each other to form a society.

Zishuo Zheng
2022/12/15