Experimental Design

Shiro Takagi

2021/6/5 -

1 Overview

My hypothesis is that

- I can transform the task-specific short-term memory to general and robust memory if I
  - encode temporal information
  - enforce agents not to use long-term memory directly
  - protect long-term memory by changing network

And, I expect that language manipulation may help to preserve the long-term memory.

I will study if this hypothesis is valid or not. To evaluate the hypothesis above, I have to

- measure the generality of the memory
- measure the robustness of the memory
- design a way to encode temporal information
- design a way to construct abstract semantics and relation by memory
- design a way to have agents to use that information
- design a way to change network to protect memory
- study the influence of the factors above on these measure
- design experiments which reflect all the requirements above

2 Generality of Memory

Our daily experience is just a set of sensory information and the reaction of our internal state to them. However, our memory (especially episodic memory) does not seem to be in such a form. Rather, it is like, say "I saw a cat yesterday". This is symbolic and far away from just a collection of sensory inputs. Although memory itself is not purely symbolic, it is tightly connected with abstract interpretation of its raw experience. I call this kind of memory general in that the memory is not unique to a specific memory.

Therefore, it is difficult to directly measure the absolute generality of a memory. Instead, I measure the generality indirectly. I hypothesize that a general memory is easier to be "used" in more experiences. In other words, it has more similarity/overlap with various experiences. We can measure "how many times a memory is used for other experiences" or "how many similar memories it has".

A further indirect measure of the generality of a memory is to measure the generalization of a reinforcement learning agents. This is because a general memory can help agents to solve more tasks. This measurement is task-specific and is limited to reinforcement learning but common and interpretable.

3 Robustness of memory

I say a memory is robust when the memory is hard to be forgotten. In hypothesis.tex, I define memory as $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$. In that sense, every memory changes even by a small bit of perturbation $\boldsymbol{\epsilon}$ since $\boldsymbol{\theta} \neq \boldsymbol{\theta} + \boldsymbol{\epsilon}$. This is not the memory we are interested in. Rather, we are interested in whether some information $f^I(\boldsymbol{\theta})$ extracted from a memory changes or not by a perturbation $\boldsymbol{\epsilon}$: we say a memory is robust when $f^I(\boldsymbol{\theta}) = f^I(\boldsymbol{\theta} + \boldsymbol{\epsilon})$.

Hence, we have to define information function $f^I : \Theta \to \mathbb{R}^d$.