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

1.1 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.

1.2 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 $f^I : \Theta \mapsto \mathbb{R}^d$. I do not introduce heuristics here. This is because I think that "useful" information depends on internal and external states. So, I consider the information function as a parametrized function. The next thing to consider is how to modify the parameters of the information function. In conventional reinforcement learning, this is done by maximizing an expected cumulative reward or some intrinsic reward. I also believe that modification is done by minimizing some energy. However, we believe that energy function is not static but dynamic in that it keeps changing by the internal and external states. I borrow the

idea from neuroscience that the brain is predictive machine. So, I decide to define the energy as the multi-scale prediction error. This belief is based on several previous works [1, 2, 3]. I consider two kind of predictions: one is the prediction on the external state and another is that on the internal state. The former tries to minimize the error between its predictive next external state and the actual next external state. The later minimizes the error between its internal state before and after the perturbation. If you consider multi-scale hierarchical architecture to the internal state, these two predictions are continuously connected. In addition to these predictions, you can also consider conventional external reward. Both predictions are functions of external state and internal states

$$f^{ex}(\boldsymbol{x}, f^I(\boldsymbol{\theta})) : \mathcal{X} \times \mathbb{R}^d \mapsto \mathcal{X}, \quad f^{in}(\boldsymbol{x}, f^I(\boldsymbol{\theta})) : \mathcal{X} \times \mathbb{R}^d \mapsto \mathbb{R}^d, \tag{1}$$

The energy is the function of prediction and the next internal/external state.

$$E^{ex}(f^{ex,t}, \boldsymbol{x}^{t+1}) : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}, \quad E^{in}(f^{in,t}, f^{in,t+1}) : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R} \tag{2}$$

Now I return to the discussion on robustness. Internal state $\boldsymbol{\theta}$ is modified so that it minimizes the energies E^{ex} and E^{in}. If this modification $\boldsymbol{\epsilon}$ satisfies $f^I(\boldsymbol{\theta}) = f^I(\boldsymbol{\theta} + \boldsymbol{\epsilon})$, we say that the memory is robust. Minimizing E^{in} can help memory to be robust but minimizing E^{ex} could hurt robustness.

1.3 Note on Robustness

I cannot control the external state (though I can to some extent by active inference [2]). I think there are three ways to install the robustness to the network $\boldsymbol{\theta}$.

The first one is changing the way to update $\boldsymbol{\theta}$. This is a mainstream in the continual learning literature [4, 5]. This approach encompasses any approach that modifies the update function $f^u(\boldsymbol{\theta}^t, \boldsymbol{\epsilon}^t) = \boldsymbol{\theta}^{t+1}$.

The second one is introducing the different structure to $\boldsymbol{\theta}$. Each element θ_i of the parameter $\boldsymbol{\theta}$ is related each other. This structure can be mathematical structure like, metric, order, etc. Also, this structure can be, say feed forward, recurrent, convolution, Hopfield and so on. This structure has influences on the robustness of the memory. The mainstream of the researches on neural network architectures study a good architecture to solve a particular task. In the similar vein, we can consider a good architecture, given a information

function f^I. For example, if the network has hidden layer, the output of the function is not differentiated by a change in $\boldsymbol{\theta}$ [6, 7].

The third one is considering the different f^I. As is evident from the definition of the robustness, whether a memory $\boldsymbol{\theta}$ is robust or not depends on f^I. Thus, even if a memory is not robust for a function f^I, it could be robust for another function $f^{I'}$. In other words, we can consider a way to extract information from $\boldsymbol{\theta}$ that is not affected so much by a change in $\boldsymbol{\theta}$.

1.4 Design Temporal Information Coding

Any biological activity is temporal. I hypothesize that preserving the temporal information enhances the robustness of the memory. This hypothesis comes from my intuition that preserving temporal information is just adding an additional dimension in the encoding space. If you preserve the information when the signal comes into, you can expect that the interference by adding memories will be mitigated.

Thus, I generalize the information and I formalize the information as follows:

$$f^{I,t} = f^{I,t}(\boldsymbol{\theta}^t, \boldsymbol{\theta}^{t-1}, ..., \boldsymbol{\theta}^0) \tag{3}$$

By construction, this is obviously more robust than non-temporal one. This is not what I am interested in.

It matters to note that the information coming into brain is continuous. Every single just a moment is a tiny fraction of instant event. What I call "event" is the accumulations of these continuous sequential signals. Thus, if I consider an event $\boldsymbol{x}$ is such millisecond scale phenomenon, we should consider a set of events $\{\boldsymbol{x}^s\}_{s=t,...,t+\tau}$ as a meaningful "event". Also, we can think that neural states encode meaningful something when we thinks it as a sequences $\{\boldsymbol{\theta}^s\}_{s=t,...,t+\tau}$. Slightly abusing the notation, we consider this information as follows:

$$f^{I,t,...,t+\tau} = f^{I,t,...,t+\tau}(\boldsymbol{\theta}^t, ..., \boldsymbol{\theta}^{t+\tau}) \tag{4}$$

Although, I think the information of a snapshot $f^{I,t}$ is not necessarily semantically the same as that of sequential states $f^{I,t,...,t+\tau}$, I use this notation for simplicity.

I will reconsider what information I represent by $f^{I,t}$ and $f^{I,t,...,t+\tau}$, respectively, after considering the remaining issues.

References

[1] Jürgen Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247, 2010.

[2] Karl Friston. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2):127–138, 2010.

[3] J Hawkings and S Blakeslee. On intelligence: How a new understanding of the brain will lead to the creation of truly intelligent machines. *An Owl Book, Henry Holt and Company, New York*, 2004.

[4] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

[5] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *arXiv preprint arXiv:1705.08690*, 2017.

[6] Sumio Watanabe. *Algebraic geometry and statistical learning theory*. Number 25. Cambridge university press, 2009.

[7] Shun-ichi Amari, Hyeyoung Park, and Tomoko Ozeki. Singularities affect dynamics of learning in neuromanifolds. *Neural computation*, 18(5):1007–1065, 2006.