Hypothesis

Shiro Takagi

2021/4/28

1 Preface

1.1 Problem setting

We consider the distributed representation. A memory of experience $\boldsymbol{x} \in X \subset \mathbb{R}^n$ is a function $f : X \mapsto \Theta$. For simplicity, we denote a memory by $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$. Note that n and p are dimension of vector $\boldsymbol{x}$ and $\boldsymbol{\theta}$.

We discuss short term memory and long term memory. Short term memory is a function $f^s : X \mapsto \Theta^s$, where $\boldsymbol{\theta}^s \in \Theta^s \subset \mathbb{R}^{p^s}$. Long term memory is a compositional function $f^l := f^c \circ f^s$, where $f^c : \Theta^s \mapsto \Theta^l$ is a memory consolidation function and $\boldsymbol{\theta}^l \in \Theta^l \subset \mathbb{R}^{p^l}$. Also, we consider memory retrieval function $f^r : \Theta^l \mapsto \Theta^s$. We assume that p^l is finite and fixed

1.2 Goal

Our goal is to find optimal memory consolidation function f^c and memory retrieval function f^r, given some criterion. Followings are what we think are desirable properties for these functions to have:

- long term memory retains short term memory's information as much as possible
- long term memory is retrievable by memory retrieval function
- memory retrieval function retrieves stored information as much as possible

In a nut shell, we want the functions such that

$$\forall \varepsilon, ||f^r(f^l(\boldsymbol{x})) - \boldsymbol{x}|| < \varepsilon. \tag{1}$$

The left hand side of the equation is just a reconstruction loss.

Another thing to consider is how to combine long term memory to short term memory. Humans seem to elegantly and naturally exploit these two memory to do a task. Thus, long term memory should be encoded and decoded such that it can be exploited easily.

1.3 Memory representation

We represent memory as just a vector with no structure $\boldsymbol{\theta}$. However, the relation between each component of a distributed representation are generally asymmetric. For example, parameters of fully connected neural network construct a hierarchical structure. Therefore, considering an optimal structure of a distributed representation, given some criterion, is another issue to consider.

Bunch of studies discussed how to construct parameters for to do tasks well (short term memory). My aim is to elucidate an optimal structure for long term memory.

1.4 Experience representation

We should also consider how to represent experience. For supervised learning, experience may be a tuple of data, loss function, and algorithm. For reinforcement learning, experience may be a tuple of state, action, and loss (reword) function. A formal definition of task by Finn et al. is helpful to consider this issue [1].

If we include loss function and algorithm in the definition of experience, it might not be suitable to call $\boldsymbol{x}$ experience and ϕ memory, because loss function and algorithm are usually included in ϕ and $\boldsymbol{x}$ is data.

We define experience at a time t as a tuple of following components:

- observation: $\boldsymbol{s}_t \in \mathcal{S}$,
- next observation: $\boldsymbol{s}_{t+1} \in \mathcal{S}$,
- action: $\boldsymbol{a}_t \in \mathcal{A}$,

where $\mathcal{S}$ is state space and $\mathcal{A}$ is action space. A state is a function from a pair of state and an action:

$$s_t : \mathcal{S} \times \mathcal{A} \mapsto \mathcal{S} \tag{2}$$

An action is a function from memory and state to action space:

$$a_t : \mathcal{S} \times \Theta \mapsto \mathcal{A} \tag{3}$$

We assume that memory consists of only short term and long term memory:

$$\Theta = \Theta^s \cup \Theta^l, \Theta^s \cap \Theta^l = \emptyset \tag{4}$$

Therefore, $\boldsymbol{x}_t = (\boldsymbol{s}_t, \boldsymbol{a}_t)$.

2 Survey

2.1 Method

I read survey papers such as [11] and pick up the literatures I think is important.

2.2 Human

Human brains is though to have complementary learning system, where hippocampus is for fast learning and neocortex is for slow learning [10]. *"More specifically, the hippocampus employs a rapid learning rate and encodes sparse representations of events to minimize interference. Conversely, the neocortex is characterized by a slow learning rate and builds overlapping representations of the learned knowledge"* [11]. It matters that "the hippocampus is thought to represent experiences in pattern separated fashion, whereby in the idealized case even highly similar events are allocated neuronal codes that are non-overlapping or orthogonal" [6].

They do not think that *"the hippocampal system receives a direct copy of the pattern of activation distributed over the higher level regions of the neocortical system; instead, the neocortical representation is thought to be re-represented in a compressed format over a much smaller number of neurons in the hippocampal system"* [10]. The point is that "such compression can often occur without loss of essential information if there is redundancy in the neocortical representations" [10] [1]. They say *"it (hippocampus) can be viewed not just as a memory store but as the teacher of the neocortical processing system"* [10]. In sum, *"The temporally extended and graded nature of retrograde amnesia would reflect the fact that information initially stored in the hippocampal memory system can become incorporated into the neocortical system only very gradually, as a result of the small size of the changes made on each reinstatement"* [10]. I think following statement crucial

> *One might then be tempted to suggest that McCloskey and Cohen simply used the wrong kind of representation and that the problem could be eliminated by using sparser patterns of activation with less overlap. However, as French (1991) has noted, reducing overlap avoids catastrophic interference at the cost of a dramatic reduction in the exploitation of shared structure — However, the existence of hippocampal amnesia, together with the sketch given earlier of the possible role of the hippocampal system in learning and memory, suggests instead that one might use the success of Rumelhart's (1990) simulation, together with the failure of McCloskey and Cohen's (1989), as the basis for understanding why there is a separate learning system in the hippocampus and why knowledge originally stored in this system is incorporated in the neocortex only gradually.* [10]

Following are particularly important answers to key questions presented in this literature

- *"The principles indicate that the hippocampus is there to provide a medium for the initial storage of memories in a form that avoids interference with the knowledge already acquired in the neocortical system"*

- *"Incorporation takes a long time to allow new knowledge to be interleaved with ongoing exposure to exemplars of the existing knowledge structure, so that eventually the new knowledge may be incorporated into the structured system already contained in the neocortex. If the changes were made rapidly, they would interfere with the system of structured knowledge built up from prior experience with other related material."* [2]

[1] Why compression? How to compress information?

[2] Intuitive but why?

Following are interesting interpretation of the relation between current artificial external memory and human memory system:

> *While parallels have been drawn between the external memory of the NTM and working memory, the characteristics of its external memory can easily be related to long-term memory systems as well. Indeed, content-based addressable external memories of this kind share functionalities with attractor networks, an architecture often used to model the computational functions performed by the CA3 subregion of the hippocampus (e.g., storage and retrieval of episodic memories). There are further points of connection between the operation of the NTM and the hippocampus: information is not stored and retained indiscriminately; instead it is selected based on an estimate of potential future relevance (see section 'Proposed Role for the Hippocampus in Circumventing the Statistics of the Environment')* [6]

An interesting observation on how mammalian memory mitigate catastrophic forgetting:

> *The distributed nature of neural coding can lead to interference between sensory and memory representations. Here, we show that the brain mitigates such interference by rotating sensory representations into orthogonal memory representations over time … The transformation of sensory information into a memory was facilitated by a combination of 'stable' neurons, which maintained their selectivity over time, and 'switching' neurons, which inverted their selectivity over time. Together, these neural responses rotated the population representation, transforming sensory inputs into memory.* [8]

2.3 Artificial neural network

Dual-weight learning system have fast-learning weight and slow-learning weight [2]. Pseudo-rehearsal does not explicitly store memory but store as probabilistic model [12]. Recent approach based on a similar idea is deep generative replay [13]. Note that this approach is inspired by the generative role of hippocampus not by neocortical function. Soltoggio et al. proposed hypothesis testing plasticity, in which confidence of consistency of cause-effect relationships determines if a memory is short-term or long-term [14]. Lopez-Paz and Ranzato proposed Gradient Episodic Memory, which impose constraint that overlap between gradient of current task and old task is sufficiently large [9]. Kamara et al. also model long term memory as generative model [4, 5] [3].

2.4 Deep Generative Dual Memory Network

Deep generative dual memory network comprises two generative models: a short-term memory (STM) and a long-term memory (LTM) [3]. What I thought are following:

- As conventions, it encodes task in memory. What should I encode? Is task the best thing to encode?
- Following Lopez-Paz et al. it measures performance with Average accuracy and Backward Transfer. Should I use them?
- They propose a balancing algorithm between old memory and new memory. But it looks heuristic. Is there another principled way to do this?
- Similarly, they propose a way to use their algorithm without task index. Is the proposed way the best way to do this?
- Benchmark task is too simple and not well-motivated.

3 Notes

Lets's reconsider the problems of current memory system. My goal is to find an answer for a good memory consolidation algorithm. To that end, thinking about how to represent memory, or, what structure memory should be is important. My thoughts are here:

- Memory should encode temporal information
- Memory should represents experience not task

[3] Is generative model the best/only way to model long term memory? What is the functionality of this model

- Ideal structure for learning/inference/task solving and that for storage and retrieval is not necessarily the same; we should separate these two requirements
- Memory is fragile but once consolidated, it is surprisingly robust
- (Conscious) memory is "used" to do something new
- Memory is associative
- Every association should be temporal

4 Hypothesis

Principle and Thoughts:

- temporal order of the memory is preserved
- exact temporal order becomes ambiguous when memory ages
- agent tries to remember the memory by first attending short term memory and then search over the long term memory
- without meaning or context, its hard to remember
- language supports memory because it makes easier to store the meaning and its relation
- for many humans, long term memory is a set of meaning and relation
- when the external signal comes in, it can re-experience the past experience
- long term memory is preserved because we usually use only the abstracted meaning and its relationships and hence do not interfere the detailed memory when remembering
- short term memory is orthogonally represented
- function from short term memory to long term memory is assumed to preserve its orthogonality and hence ideally we can remember all the past experience with hight accuracy

- We just cannot remember because we usually use the abstracted meaning and relation of the memory

- it is hard to reconstruct the detailed memory because abstracted meaning and relation are associated with tons of experiences, other meanings, and relations and remembering the abstract ones are by far easier

- external signal is very informative in that it literally exactly reproduce the past experience, making it easier to remember the details

I hypothesize that

- we can remember the detailed memories but find it hard to remember them because the detailed memory is temporally complicated and easily interfered by "easy" memory like abstract meaning and relation, and we usually use only the meaning and relation

- function from short term memory to long term memory is temporal and its temporal index plays a roll of key of the memory?

- a single episode (memory) is not compressed. Compressed memory is the result of interference and use of abstract memory

5 Hypothesis

Question

- How to transform the task-specific short-term memory to general and robust long-term memory?

Response

- Long-term memory itself is not general but specific

- Long-term memory is robust because it encodes temporal information and is encoded in a temporal manner

- Long-term memory is robust because we usually use abstract memory like meaning and relations that is emerged from several long-term memories

- (Long-term memory is robust because of gene expression)
- Almost all sensory memory is forgotten and almost all short-term memory is forgotten
    - Long-term memory is retained in the way different from that for encoding memory at first

Hypothesis

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

A natural consequence is that language is complementary to long-term memory. Language manipulation may help to preserve long-term memory.

6 Survey for hypothesis refinement

I find the idea in [7] close to the idea above in that they propose to attend to an aggregated chunks of memories not to attend to the raw memory. I think it matters to consider

- a way to construct a chunk
    - They propose to divide temporally
    - I thinks that chunk is not only temporal but also semantic
    - Temporal chunk may not be uniform: The size of chunk differs, depending on when the episode occurs
- a motivation for agents to use small chunks him/herself
    - Ideally agents should learn/find to do so
- a way to transform episode to long-term memory
    - They restore the memory in a raw form

7 Breakdown Hypothesis

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

7.1 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.

7.2 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 [**?**, **?**, **?**]. 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

$$f^{ex}(\boldsymbol{x}, f^I(\boldsymbol{\theta})) : \mathcal{X} \times \mathbb{R}^d \mapsto \mathcal{X}, \quad f^{in}(\boldsymbol{x}, f^I(\boldsymbol{\theta})) : \mathcal{X} \times \mathbb{R}^d \mapsto \mathbb{R}^d, \tag{5}$$

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

$$E^{ex}(f^{ex,t}, \boldsymbol{x}^{t+1}) : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}, \quad E^{in}(f^{in,t}, f^{in,t+1}) : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R} \tag{6}$$

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.

7.3 Note on Robustness

I cannot control the external state (though I can to some extent by active inference [**?**]). 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 [**?**, 13]. 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 $\boldsymbol{\theta}$ [**?**, **?**].

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 $\boldsymbol{\theta}$.

7.4 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:

$$f^{I,t} = f^{I,t}(\boldsymbol{\theta}^t, \boldsymbol{\theta}^{t-1}, ..., \boldsymbol{\theta}^0) \tag{7}$$

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:

$$f^{I,t,...,t+\tau} = f^{I,t,...,t+\tau}(\boldsymbol{\theta}^t, ..., \boldsymbol{\theta}^{t+\tau}) \tag{8}$$

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.

7.5 Design Semantics/Relation Representation

Human understand a notion of "research" and know that it includes "science", for example. Or, we can use a more abstract notion of "object" and apply an "operation" on it. Human beings seem to excel at these symbolic manipulation. I believe that memory is "used" for these operations. Abstract semantics is subset of information f^I: f^{sem}. The relation between the semantics is a function of multiple semantics: $f^{rel} = f^{rel}(f^I \times ... \times f^{I'})$.

I support the view that the semantics of neural representation is designed through the culture the agent is in [**?**]. Repeated activations of neurons constructs an abstract concept and the symbols are attached to these concepts and segments the neural representation space. Thus, the abstract concept is formed through energy minimization above and semantics is grounded to these abstract through social interactions. I hypothesize that human attach the symbol because it improves the predictability of internal and external states. In other words, we do so because it is useful. I think that the relation is also formed through this process.

With regards to the operation, I think that symbolic operation is a generalization of the physical action in neural representation space. When we move a cup, the cup will change it position after the action. In the similar vein, we act on a symbol and produce another symbol. Planning, making a hypothesis, proving a proposition, all of these mental actions can be regarded as like this. I find a book that presents a similar idea a bit [**?**]. I also think that symbolic operation occurs just because it is "useful".

7.6 Design information usage scheme

I said agents use semantics and relations because it is "useful". So, I have to clarify what I mean by "useful". What is the situation where agents do not directly use the raw memory $\boldsymbol{\theta}$ but uses its information $f^I(\boldsymbol{\theta})$. If the goal is not remembering or dealing with a particular situation, agents do not have to remember the experience the same as that when it was stored. If agents have to face various situation, it is better to use the abstract representation as well. Also, when agents are required to quickly adapt to the situation and extracting the raw memory takes cost, using accumulated information could be better.

7.7 Design structural change for memory protection

Neuron strengthens its synaptic connection by two ways: chemical reaction and gene expression. The former is needed for the short term memory and the later is needed for the long term memory. I do not think that I have to design a biologically plausible memory mechanism for my purpose but I could use multiple mechanisms to form a memory. In the brain, new synapse is generated if the information is worth storing as a long-term memory. But increasing synapses infinitely is not plausible. Natural solution to this will be the occasional pruning unnecessary synapse. I still have no idea on what is the optimal adding/pruning ratio, timing, and amount. The efficiency of this scheme tightly depends on how a neuron assembly represents memories.

8 Rethinking hypotheses

I found it messy to tackle all the hypotheses I proposed above at the same time. Thus, I will rethink the hypothesis and define the hypothesis I will focus on.

References

[1] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.

[2] Geoffrey E Hinton and David C Plaut. Using fast weights to deblur old memories. In *Proceedings of the ninth annual conference of the Cognitive Science Society*, pages 177–186, 1987.

[3] Nitin Kamra, Umang Gupta, and Yan Liu. Deep generative dual memory network for continual learning. *arXiv preprint arXiv:1710.10368*, 2017.

[4] Ronald Kemker and Christopher Kanan. Fearnet: Brain-inspired model for incremental learning. *arXiv preprint arXiv:1711.10563*, 2017.

[5] Ronald Kemker, Marc McClure, Angelina Abitino, Tyler Hayes, and Christopher Kanan. Measuring catastrophic forgetting in neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[6] Dharshan Kumaran, Demis Hassabis, and James L McClelland. What learning systems do intelligent agents need? complementary learning systems theory updated. *Trends in cognitive sciences*, 20(7):512–534, 2016.

[7] Andrew K. Lampinen, Stephanie C. Y. Chan, Andrea Banino, and Felix Hill. Towards mental time travel: a hierarchical memory for reinforcement learning agents. *arXiv preprint arXiv:2105.14039*, 2021.

[8] Alexandra Libby and Timothy J Buschman. Rotational dynamics reduce interference between sensory and memory representations. *Nature Neuroscience*, pages 1–12, 2021.

[9] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *arXiv preprint arXiv:1706.08840*, 2017.

[10] James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419, 1995.

[11] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

[12] Anthony Robins. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2):123–146, 1995.

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

[14] Andrea Soltoggio and Frank van der Velde. Neural plasticity for rich and uncertain robotic information streams. *Frontiers in neurorobotics*, 9:12, 2015.