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 [8] 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 [7]. *"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"* [8]. 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" [5].

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"* [7]. The point is that "such compression can often occur without loss of essential information if there is redundancy in the neocortical representations" [7] [1]. They say *"it (hippocampus) can be viewed not just as a memory store but as the teacher of the neocortical processing system"* [7]. 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"* [7]. 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.* [7]

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')* [5]

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 [9]. Recent approach based on a similar idea is deep generative replay [10]. 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 [11]. Lopez-Paz and Ranzato proposed Gradient Episodic Memory, which impose constraint that overlap between gradient of current task and old task is sufficiently large [6]. Kamara et al. also model long term memory as generative model [3, 4] [3].

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.

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

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

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

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

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

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