Contextual Representation Ensembling

Tyler M. Tomita (ttomita@jhu.edu)
Department of Psychological and Brain Sciences
Johns Hopkins University
3400 N. Charles Street, Baltimore, MD 21218

Abstract

Real-world agents must be able to efficiently acquire new skills over a lifetime, a process called "continual learning." Current continual machine learning models fall short because they do not selectively and flexibly transfer prior knowledge to novel contexts. We propose a cognitively-inspired model called Contextual Representation Ensembling (CRE), which fills this gap. We compared CRE to other state-of-the-art continual machine learning models as well as other baseline models on a simulated continual learning experiment. CRE demonstrated superior transfer to novel contexts and superior remembering when old contexts are re-encountered. Our results suggest that, in order to achieve efficient continual learning in the real world, an agent must have two abilities: (i) they must be able to recognize context cues within the environment in order to infer what prior knowledge might be relevant to the current context and (ii) they must be able to flexibly recombine prior knowledge.

Keywords: Continual learning, task sets, rule-switching, transfer, catastrophic forgetting

Introduction

Humans are capable of "continual learning": we can use what we have learned in one context to quickly learn something in a new context, progressively building and refining a repertoire of skills over of a lifetime. Traditionally, machine learning systems would struggle with learning a sequence of tasks, because learning a new task produced "catastrophic forgetting" of what had been learned before. While there are now both algorithmic and architectural solutions that minimize catastrophic forgetting in simple settings, machines still substantially forget when faced with real-world sequences of tasks. Furthermore, machine learning systems still cannot flexibly and rapidly re-use knowledge as humans can.

Here we present a novel artificial neural network (ANN) architecture and training procedure for learning a sequence of supervised learning tasks, guided by an overarching hypothesis about how continual learning is achieved in humans: we propose that humans learn and store multiple knowledge representations for any input, and solve tasks by combining ("ensembling") knowledge representations that are appropriate to the current context. Recognizing the relevant context aids performance, not only because it enables us to retrieve relevant knowledge, but also because our new learning will not interfere with context-irrelevant knowledge, which is stored in other representations. We call this strategy for continual learning *Contextual Representation Ensembling* (CRE).

Contextual Representation Ensembling

The CRE architecture (Fig 1) is based on a mixture of experts (MoE) design, and resembles the cognitively-inspired DynaMoE model for continual machine learning. Like DynaMoE, our CRE model starts with a single "expert" module, and dynamically recruits and trains additional experts as new tasks require them. In contrast to DynaMoE, which can only reuse individual experts, our model makes use of an "Ensembler" module, which enables representations of multiple experts to be flexibly recombined to solve novel tasks. We call these ensembles of experts "schemas". Furthermore, unlike DynaMoE, our model makes use of a context recognizer module, which exploits task-informative context cues within the input. The context recognizer relies on an episodic memory bank, which stores examples of previously encountered contexts and the optimal schemas associated with them. Recognizing context from environmental cues helps to (i) initialize a compact set of existing schemas that may be relevant in a novel but related context is encountered (via nearest neighbors matching of the novel context to prior contexts in episodic memory) and (ii) immediately reinstate the context-appropriate schema when an old context is re-encountered.

Results

We simulated a series of four binary classification tasks (Fig 2a), and trained several ANN algorithms on them. The probability distribution of the input is identical in all tasks. However, the classification rule differs between the tasks. The first 17 dimensions of the input $X \in \mathbb{R}^{20}$ is uniformly distributed over $[1,-1]$. The last three dimensions are configured such that they have unique distributions for each task. Thus, these dimensions are task-informative "context cues." In Task 1, the class label Y is determined by the XOR function on the first two dimensions:

$$Y = \text{XOR}(X_1, X_2) = \begin{cases} 1 & \text{if } X_1 * X_2 > 1 \\ 0, & \text{otherwise} \end{cases}$$

Similarly, the class labels in Tasks 2 and 3 are determined by $\text{XOR}(X_3, X_4)$ and $\text{XOR}(X_5, X_6)$, respectively. In Task 4, $Y = 1$ if $X_1 * X_2 + X_3 * X_4 > 0$ and $Y = 0$ otherwise. Because Tasks 1 and 2 require learning representations that multiply $X_1 * X_2$ and $X_3 * X_4$, respectively, Task 4 can be learned much more quickly by selectively transferring these two representations and linearly combining them.

We compared CRE to four other algorithms: 1) DynaMoE, 2) Elastic Weight Consolidation (EWC), 3) a basic ANN that

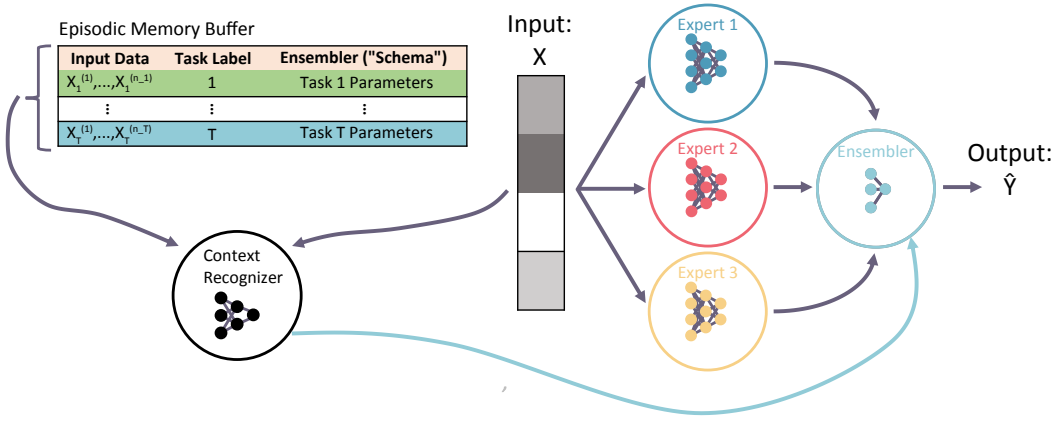

Figure 1: The CRE architecture.

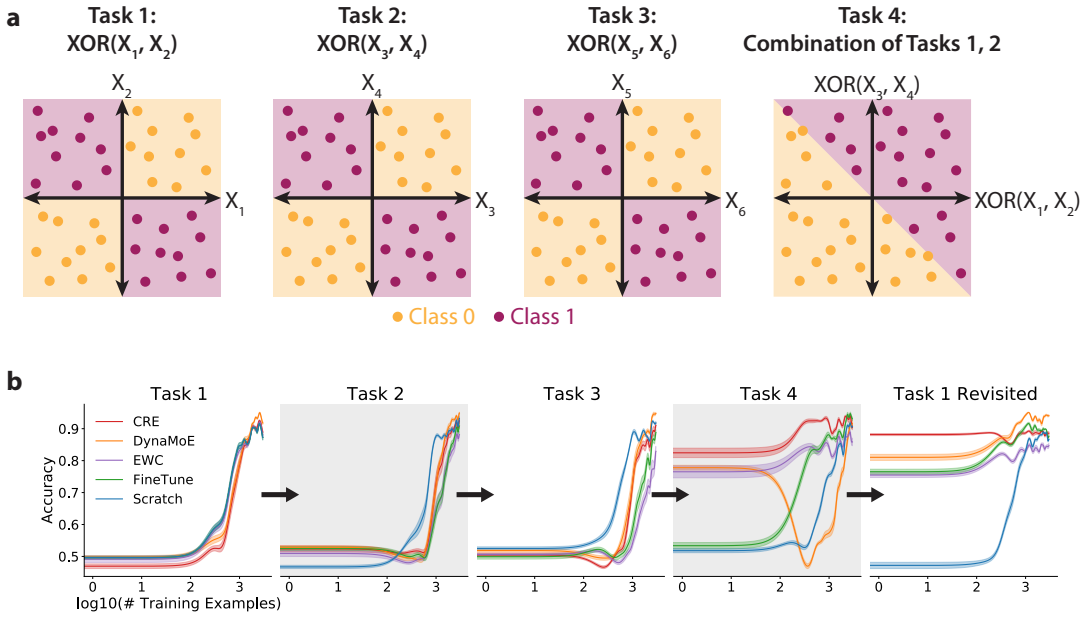

Figure 2: (a): Visualization of the four binary classification tasks in input space.(b): Classification accuracy comparison of the different algorithms as they sequentially proceed to learn Tasks 1-4 in order followed by revisiting Task 1.

fine tunes on each subsequent task (FineTune), and 4) a basic ANN that trains from scratch on each task (Scratch). All algorithms were trained incrementally in sequential order from Tasks 1-4 followed by revisiting Task 1. We found that CRE performed comparably to all other algorithms on Tasks 1-3. However, CRE exhibited superior transfer in Task 4 and superior memory and recognition of Task 1 when it was re-encountered.

Conclusions

Real-world agents must learn rapidly in novel context by reusing and recombining prior knowledge; the inability to do so can have dire consequences In real-world scenarios in which feedback is delayed or sparse, or in which just one or two mistakes is too costly, rapid performance is critical. We demonstrated that CRE enables rapid performance in two ways. First, by implementing an ensembler, prior knowledge ("experts") can be flexibly recombined to rapidly learn novel tasks. Second, recognizing old contexts enables immediate remembering, and recognizing similar novel contexts enables rapid inference of what prior knowledge might be relevant.

References

Tsuda, B., Tye, K. M., Siegelmann, H. T., & Sejnowski, T. J. (2020). A modeling framework for adaptive lifelong learning with transfer and savings through gating in the prefrontal cortex. Proceedings of the National Academy of Sciences, 117(47), 29872-29882.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., ... & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences, 114(13), 3521-3526.