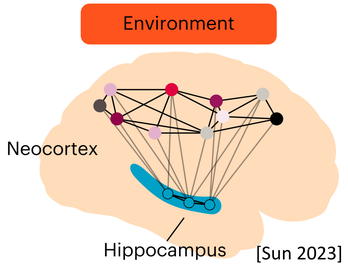
Remembering Transformer for Continual Learning

Yuwei Sun, Ippei Fujisawa, Arthur Juliani, Jun Sakuma, Ryota Kanai

Catastrophic Forgetting

- Catastrophic Forgetting (CF): new task knowledge interferes with old knowledge, causing previously learned tasks forgotten.
- Existing fine-tuning and regularization methods necessitate task identity information and cannot eliminate task interference, while soft parameter sharing encounters an increasing parameter size.

Complementary Learning Systems Theory

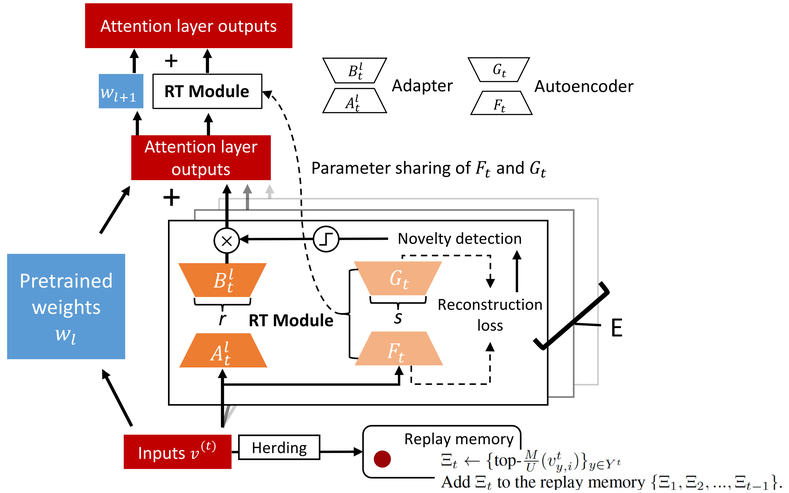


[Sun 2023]

- Hippocampus rapidly encodes task data and consolidates the task knowledge into the Neocortex by forming new neural connections.
- Hippocampus developed a novelty detection mechanism to facilitate consolidation by switching among neural modules for various tasks.

Remembering Transformer

Key idea: leverage the mixture-of-adapters that are sparsely activated with a novelty detection mechanism in a pretrained Transformer.

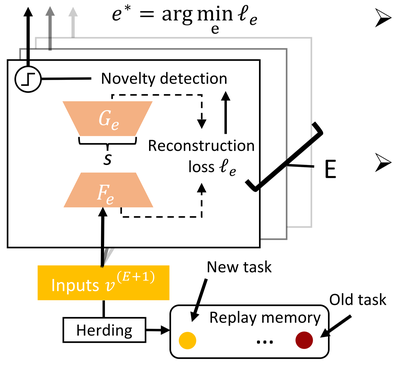


Mixture-of-Adapters in ViT

Low-Rank Adaptation: $W_l + \Delta W_l = W_l + B^l A^l$, where $A^l \in \mathbb{R}^{r\times D}, B^l \in \mathbb{R}^{D\times r}$

Mixture-of-Adapters: $W_l v_{l-1} + \sum_{e=1}^{E} W_g(e) B^l_e A^l_e v_{l-1}$

Generative Routing



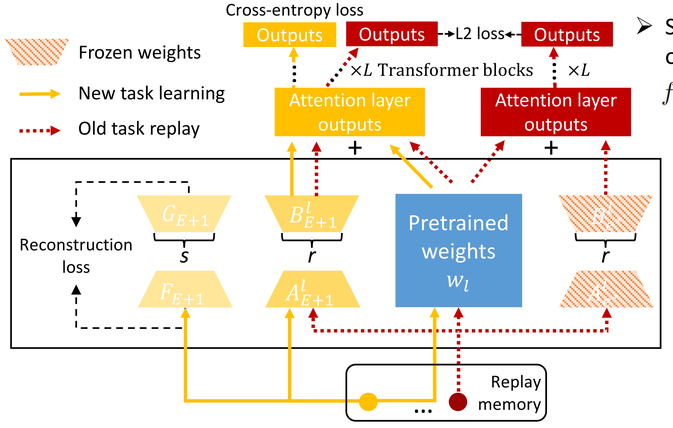
- During inference, the autoencoders, each encoding different task knowledge, predict routing weights to allocate a sample to the most relevant adapter.
- The reconstruction loss represents the similarity between the current task and the old knowledge encoded in these autoencoders.

$$\ell_e = (\sigma(v) - \hat{G}_e\hat{F}_e\sigma(v))^2,\ e^* = \arg\min_e \ell_e,$$

$$W_g(e) = \begin{cases} 1 & \text{if } e = e^* \\ 0 & \text{otherwise.} \end{cases}$$

Knowledge Distillation

We explore a scenario where the number of adapters is constrained and then propose the adapter fusion to identify and aggregate resembling adapters.

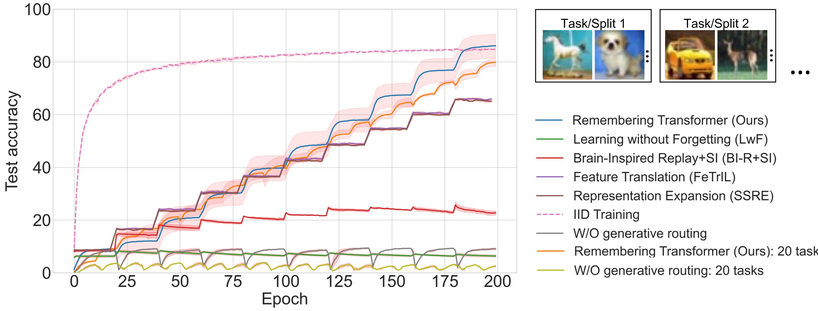


- Soft probability output of the old adapter $f(\Xi_{e^*}; \{\hat{\theta}_{\text{ViT}}, \hat{\theta}^{e^*}_{\text{adapter}}\})$
- The old adapter is removed, and its task samples are distributed to the newly learned adapter *E+1*.

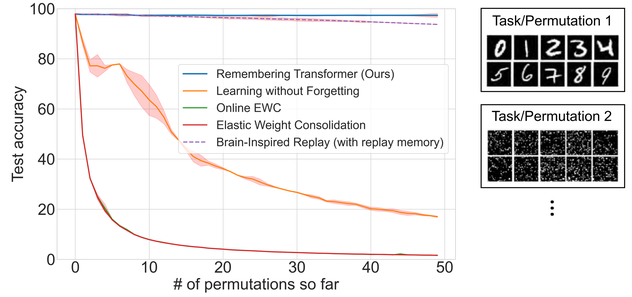
$$\text{Gate}(e^*) \leftarrow E + 1$$

Continual Learning Tasks

- Split task: a dataset is split into several subsets of equal numbers of classes, with each subset as a task.



- Permutation task: the input pixels of an image are shuffled with a random permutation for each task.



Model Efficacy Under a Constrained Capacity

#Adapters (memory footprint)	Test accuracy (%)
5 (0.37M)	**99.3 ± 0.24**
3 (0.22M)	93.2 ± 0.72
2 (0.15M)	87.1 ± 0.85

CLOM	88.0 ± 0.48
SSRE	90.5 ± 0.61
FeTrIL	90.9 ± 0.38

Remembering Transformer demonstrated SOTA continual learning *task accuracy* and *parameter efficiency* through the mixture-of-adapters and generative routing in ViT.

arXiv:2404.07518
yuwei_sun@araya.org