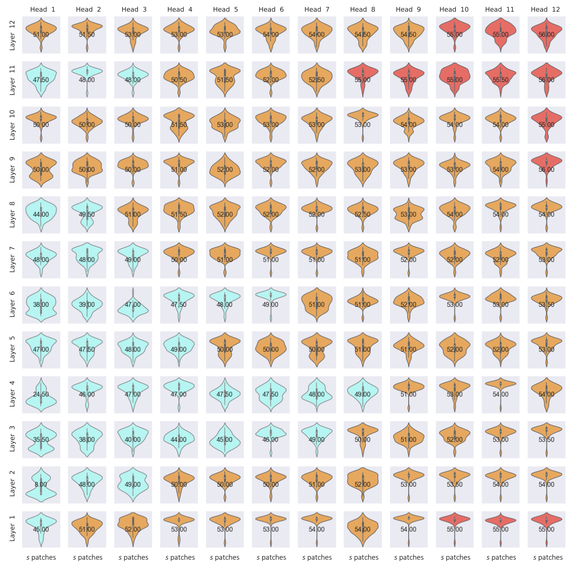
Associative Transformer Is A Sparse Representation Learner

Yuwei Sun[1,2], Hideya Ochiai[1], Zhirong Wu[3], Stephen Lin[3], Ryota Kanai[4]

[1]The University of Tokyo, [2]RIKEN, [3]Microsoft Research, [4]Araya

INTRODUCTION

- Transformer models use pairwise attention to establish correlations among disparate segments of input information.
- Competition, which reveals naturally sparser interactions among attention heads in pairwise attention, is important for learning meaningful representations.



The sparsity of attention for a specific patch's interactions with other patches is computed as $\arg\min_s \sum_{j=1}^{s} A^{i,j} \geq 0.9$.

Unlike convolution operations in CNNs, self-attention in Transformers does not possess inductive biases that allow it to attend to different segments of the input data. **We aim to introduce architecture that can foster competition among patches by constraining the number of patches that each head can focus on, thereby inducing an inductive bias for meaningful patch learning.**

METHODS

We propose the Associative Transformer (AiT) with a novel *global workspace layer* building upon recent neuroscience studies of the Global Workspace Theory and associative memory.

Modularization of knowledge can find resonance with the neuroscientific grounding of the Global Workspace Theory (GWT). GWT explains a cognitive architecture where diverse specialized modules compete to write information into shared workspace through a communication bottleneck.

When examining information retrieval in the human brain, it is evident that memory typically encompasses both working memory and long-term memory in the hippocampus. Specifically, the hippocampus operates on Hebbian learning, akin to the associative memory found in modern Hopfield networks.

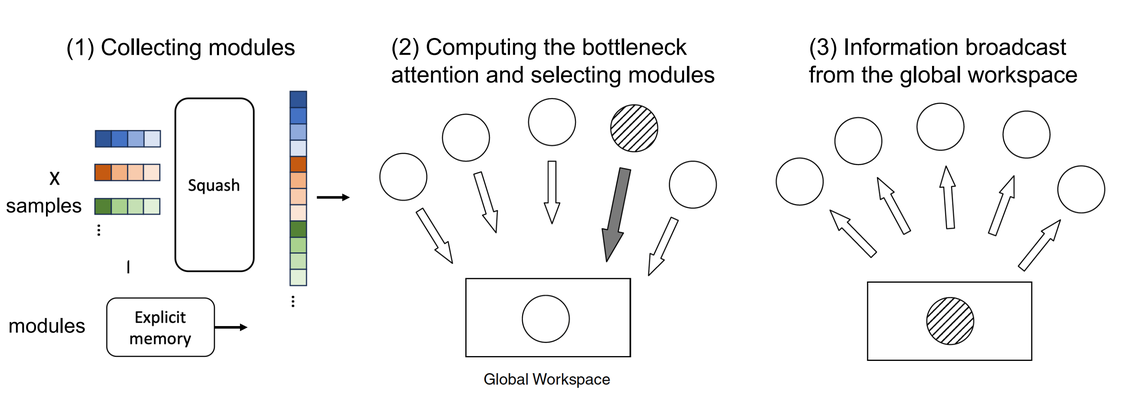


Figure 2. Inducing global workspace for emerging module specialization.

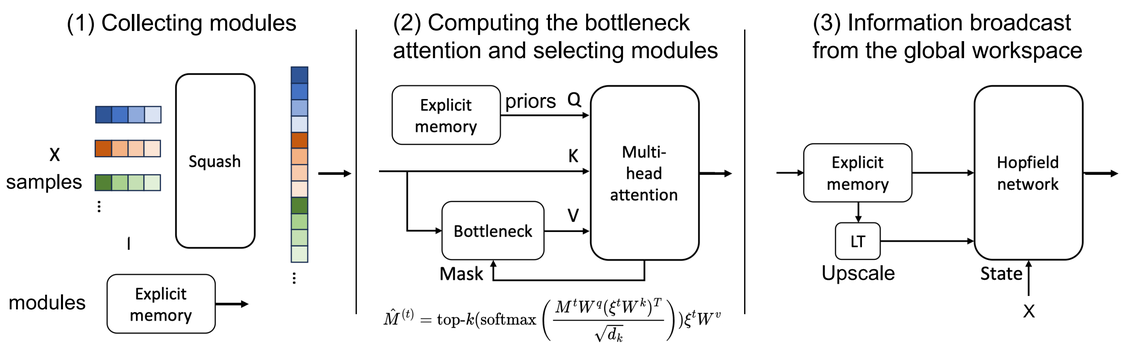


Figure 3. The scheme of the global workspace layer.

- The squash layer concatenates patches within one batch $V \in \mathbb{R}^{B\times N\times E}$ into vectors $V \in \mathbb{R}^{(B\times N)\times E}$.
- Squashed representations are projected to a low-rank latent space of dimension $D << E$ and then are sparsely selected and stored in the explicit memory via a fixed bottleneck $k << (B \times N)$.
- Explicit low-rank memory with limited slots learns M *priors* $\gamma = \mathbb{R}^{M\times D}$, to compute the bottleneck attentions that extract different sets of patches from the input.

$$\text{head}_i^t = \text{top-}k(\text{A}_i^t)\Xi^t W_t^V,\ \hat{\gamma}^t = \text{LN}(\text{Concat}(\text{head}_1^t, \ldots, \text{head}_A^t)W^O).$$

- The Hopfield network utilizes the memory to reconstruct the input, where a learnable linear transformation (LT) scales the memory contents $f_{\text{LT}}(\gamma^{t+1})$ to match the input dimension E.

$$E(\xi^t) = -\text{lse}(\beta, f_{\text{LT}}(\gamma^{t+1})\xi^t) + \frac{1}{2}\xi^t\xi^{tT} + \beta^{-1}\log M + \frac{1}{2}\zeta^2,\ \hat{\xi}^t = \arg\min_{\xi^t} E(\xi^t).$$

Bottleneck Attention Balance Loss

Learning specialized priors in layers cascaded in depth requires a mechanism that counteracts the inherent loss of input specificity, as information flows through multiple layers.

The bottleneck attention balance loss encourages the selection of diverse patches from different input positions into the shared workspace.

Accumulative attention scores: $\ell_{\text{importance}_{i,l}} = \sum_{j=1}^{M} \text{A}_{i,j,l}^t,$

Chosen instances: $\ell_{\text{loads}_{i,l}} = \sum_{j=1}^{M} (\text{A}_{i,j,l}^t > 0),$

$$\ell_{\text{bottleneck}_i} = \frac{\text{Var}(\{\ell_{\text{importance}_{i,l}}\}_{l=1}^{B\times N})}{(\frac{1}{B\times N}\sum_{l=1}^{B\times N}\ell_{\text{importance}_{i,l}})^2 + \epsilon} + \frac{\text{Var}(\{\ell_{\text{loads}_{i,l}}\}_{l=1}^{B\times N})}{(\frac{1}{B\times N}\sum_{l=1}^{B\times N}\ell_{\text{loads}_{i,l}})^2 + \epsilon}.$$

RESULTS

Our study demonstrates that AiT outperforms the previously employed attention-based approaches such as the Coordination, the Set Transformer, and the Perceiver when applied to vision-related tasks.

Table 1. Model performance comparison in image classification tasks

Methods	CIFAR10	CIFAR100	Triangle	Average	Model Size
AiT-Base (Ours)	**85.44**	**59.10**	99.59	**81.38**	91.0
AiT-Small (Ours)	83.34	56.30	99.47	79.70	15.8
Coordination	75.31	43.90	91.66	70.29	2.2
Coordination-DH	72.49	51.70	81.78	68.66	16.6
Coordination-D	74.50	40.69	86.28	67.16	2.2
Coordination-H	78.51	48.59	72.53	66.54	8.4
ViT-Base	83.82	57.92	**99.63**	80.46	85.7
ViT-Small	79.53	53.19	99.47	77.40	14.9
Perceiver	82.52	52.64	96.78	77.31	44.9
Set Transformer	73.42	40.19	60.31	57.97	2.2
BRIMs	60.10	31.75	-	45.93	4.4
Luna	47.86	23.38	-	35.62	77.6

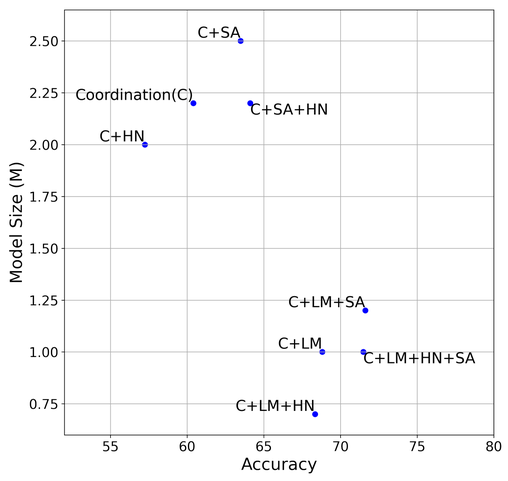


Figure 4. Model size vs. accuracy.

Using the low-rank memory (LM) that has a more diverse set of priors showed benefits in both improving the performance and decreasing the model size. The Hopfield network (HN) maintained the model performance while reducing the model size by replacing the cross-attention with more efficient retrieval, which was effective only when either the LM or SA component was applied. We assume that the retrieval relies on a diverse set of priors, which is enabled using the enhanced bottleneck attention and the learning through self-attention.

- Increased Bottleneck Attention Distribution Sparsity

Figure 5. Selected patches by the bottleneck attention.

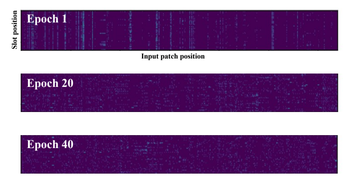
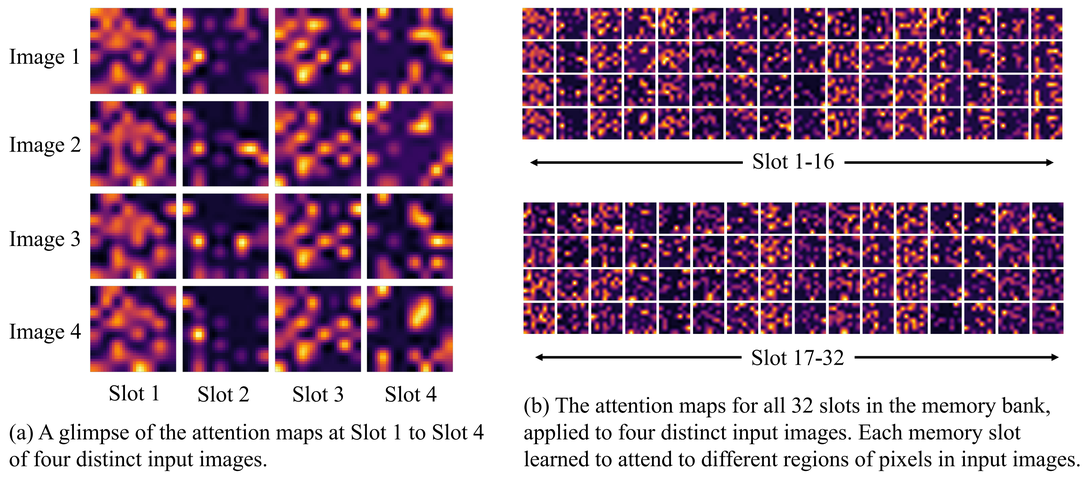


Figure 6. Selecting diverse patches from different input positions.

- Prior Specialization: patches in one image can be attended sparsely by different priors through the bottleneck attention.



(a) A glimpse of the attention maps at Slot 1 to Slot 4 of four distinct input images.

(b) The attention maps for all 32 slots in the memory bank, applied to four distinct input images. Each memory slot learned to attend to different regions of pixels in input images.

Figure 7. Learned bottleneck attention maps in CIFAR10.

CONCLUSIONS

Associative Transformer (AiT) leverages a diverse set of priors with the emerging specialization property to enable sparse association among representations via the Hopfield network. The comprehensive experiments demonstrate AiT's efficacy compared to conventional models including the previous coordination method.

Corresponding author: ywsun@g.ecc.u-tokyo.ac.jp

arXiv:2309.12862