(CVPR 2025)
Associative Transformer

November 2, 2025

Seoul National University

Intro

Transformers rely on *pairwise self-attention*, correlating every token with every other.

- computationally expensive
- biologically implausible - it lacks localized and contextual learning like the human brain.

Previous *sparse attention* models such as the Coordination method tried to use a bottleneck mechanism to restrict attention to key tokens, but they suffered from

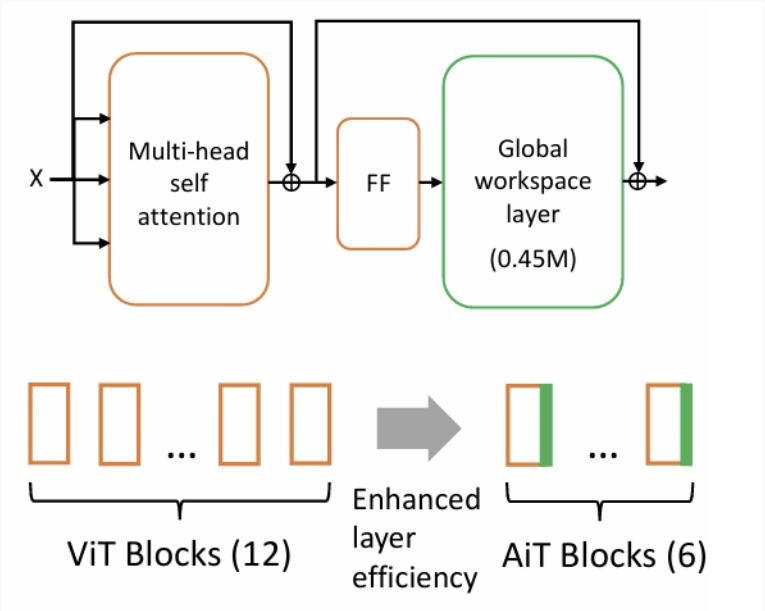
- parameter inefficiency
- poor performance on complex relational reasoning tasks

Intro

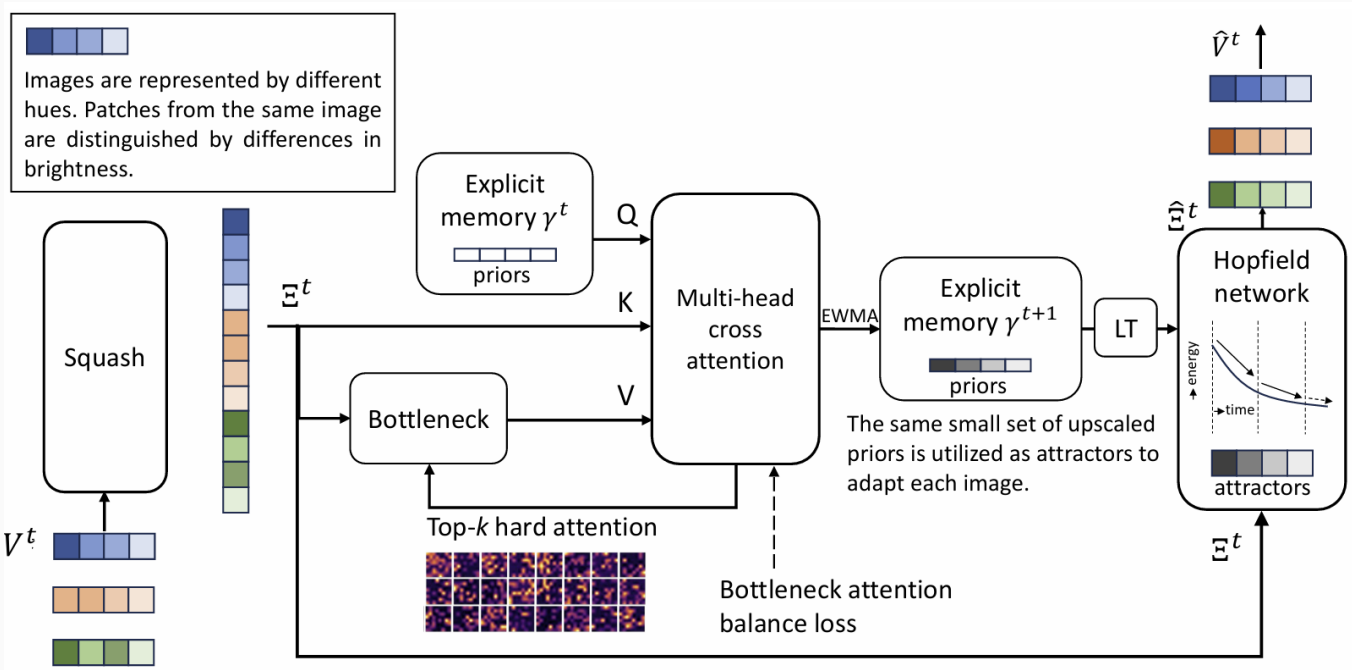
This paper introduces the Associative Transformer (AiT) to overcome these problems.

Ait integrates:

1. a low-rank explicit memory for learning diverse local pattern.
2. an associative Hopfield memory for reconstructing token representations.
3. achieving higher efficiency and accuracy with fewer parameters.



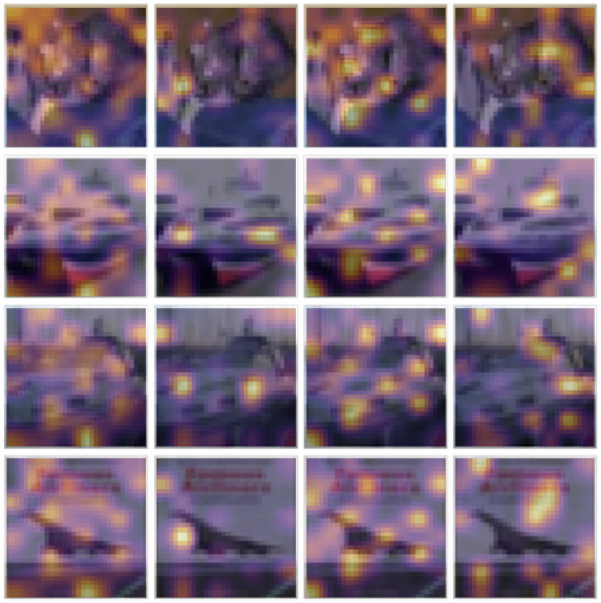
The scheme of the Associative Transformer (AiT).



(a) Global workspace layer

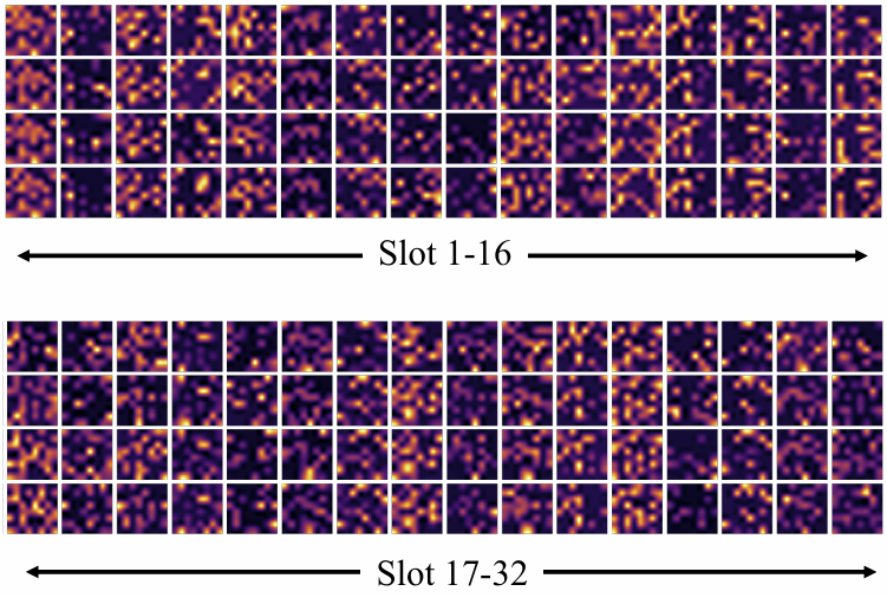
Figure 1: (a) In a global workspace layer, the input $R^{B\times N\times E}$ is squashed into vectors $R^{BN\times E}$. The squashed representations are projected to a latent space of dimension $D << E$ and are sparsely selected to update the explicit memory via a fixed bottleneck $k << BN$. The Hopfield network utilizes the memory to reconstruct the input tokens, where a learnable linear transformation (LT) scales the memory contetnts back to the input dimension E.

Prior-guided attention visualization



Slot 1 Slot 2 Slot 3 Slot 4

(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.

Experiments-classification task

Methods	CIFAR10 (%)	CIFAR100 (%)	Triangle (%)	Average (%)	Size (M)	#FLOPs
AiT-Base	**85.44 ± 0.31**	**60.78 ± 0.25**	**99.64 ± 0.14**	**81.95**	91.0	5.77×10^9
AiT-Medium	84.59 ± 0.27	60.58 ± 0.32	99.57 ± 0.16	81.58	45.9	2.89×10^9
AiT-Small	83.34 ± 0.44	56.30 ± 0.38	99.47 ± 0.09	79.70	15.8	9.64×10^8
Coordination [13]	75.31 ± 0.72	43.90 ± 0.21	91.66 ± 0.56	70.29	2.2	1.46×10^8
Coordination-DH [13]	72.49 ± 0.61	51.70 ± 0.75	81.78 ± 0.59	68.66	16.6	3.15×10^8
Coordination-D [13]	74.50 ± 0.68	40.69 ± 0.39	86.28 ± 0.73	67.16	2.2	2.91×10^8
Coordination-H [13]	78.51 ± 0.58	48.59 ± 0.72	72.53 ± 0.35	66.54	8.4	1.59×10^8
ViT-Base [2]	83.82 ± 0.17	57.92 ± 0.40	99.63 ± 0.15	80.46	85.7	5.60×10^9
ViT-Medium [2]	82.41 ± 0.11	55.78 ± 0.09	99.62 ± 0.04	79.27	42.7	2.81×10^9
ViT-Small [2]	79.53 ± 0.36	53.19 ± 0.37	99.47 ± 0.07	77.40	14.9	9.36×10^8
Perceiver [26]	82.52 ± 0.82	52.64 ± 0.44	96.78 ± 0.32	77.31	44.9	2.37×10^9
Set Transformer [23]	73.42 ± 0.43	40.19 ± 0.53	60.31 ± 0.29	57.97	2.2	1.11×10^8
BRIMs [39]	60.10 ± 0.50	31.75 ± 0.28	58.34 ± 0.43	50.06	4.4	1.43×10^8
Luna [25]	47.86 ± 0.53	23.38 ± 0.06	57.26 ± 0.19	42.83	77.6	5.08×10^9

Table 1. Performance comparison in the classification tasks.

Experiments-importance of Bottleneck

Models	CIFAR10 (%)	CIFAR100 (%)	Triangle (%)	Average (%)
AiT	**83.34 ± 0.44**	**56.30 ± 0.38**	**99.47 ± 0.09**	**79.70**
Reset Memory	81.94 ± 0.41	55.96 ± 0.39	99.46 ± 0.04	79.12
W/O Attention Balance Loss	81.89 ± 0.39	54.72 ± 0.33	99.44 ± 0.08	78.68
W/O Hopfield	81.03 ± 0.45	54.96 ± 0.28	99.44 ± 0.03	78.48
W/O Memory	79.53 ± 0.36	53.19 ± 0.37	99.47 ± 0.07	77.40
W/O Bottleneck	75.40 ± 0.48	46.53 ± 0.41	93.33 ± 0.15	73.75
W/O SA	72.72 ± 0.57	47.75 ± 0.31	99.46 ± 0.04	73.31

Table 3. The ablation study demonstrated that leveraging all the components resulted in the best performance.

Experiments-parameter efficiency

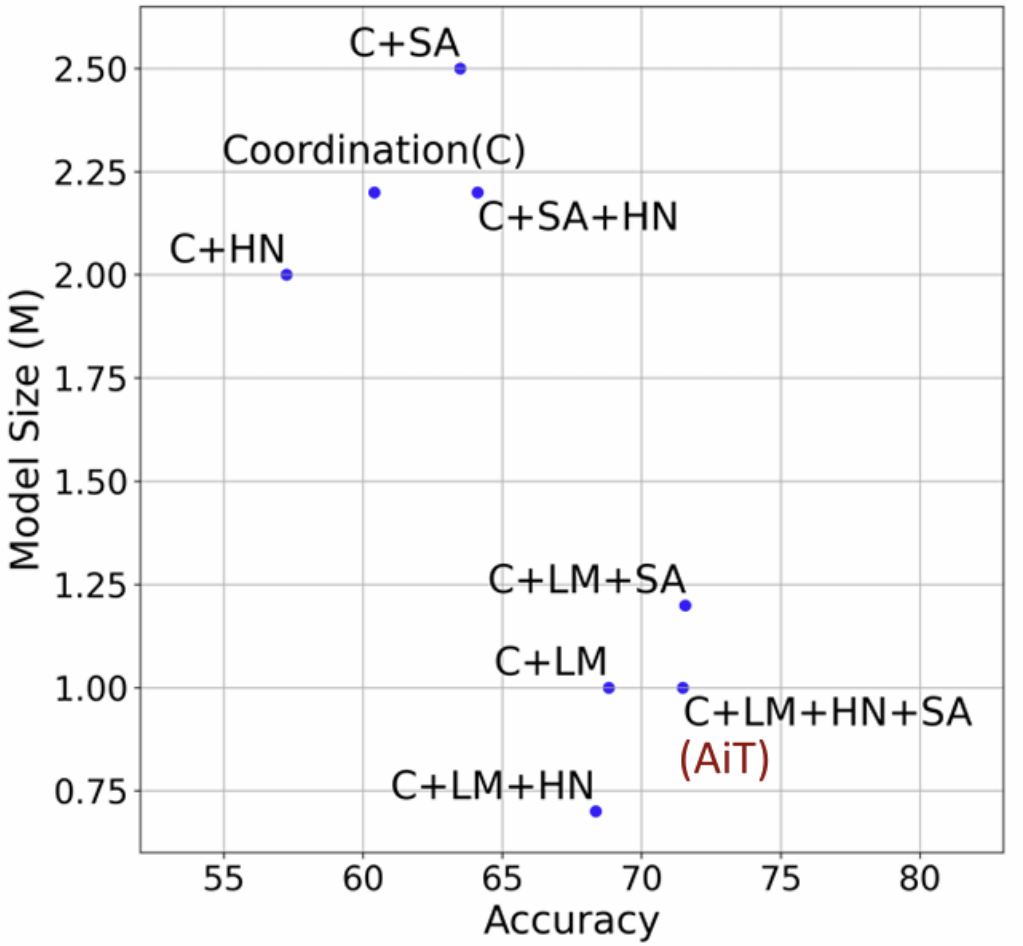


Figure 3. Model size vs. test accuracy for various model configurations. Consolidating all the components in the Coordination block resulted in the best performance of 71.49% maintaining a compact model size of 1.0M.

Algorithms

Algorithm 1 Global Workspace Layer

Main program

Input: tokens from the previous layers: $V \in \mathbb{R}^{B\times N\times E}$; learnable low-rank priors: $\gamma \in \mathbb{R}^{M\times D}$

Squash layer concatenates tokens in the batch: $\Xi \in \mathbb{R}^{BN\times E} \leftarrow V \in \mathbb{R}^{B\times N\times E}$

Project Ξ to a latent space with a dimension $D \ll E$: $\Xi W_i^K \in \mathbb{R}^{BN\times D}$

Obtain the attention scores over the projected tokens using priors γ: $\mathrm{A}_i(\gamma, \Xi) = \mathrm{softmax}(\frac{\gamma(\Xi W_i^K)^T}{\sqrt{D}})$

Tokens compete to write in memory via a bottleneck with capacity k: $h_i = \text{top-}k(\mathrm{A}_i)\Xi W^V$ (*Bottleneck Attention Balance Loss is employed for a more diverse token selection)

Update priors γ^t with Exponentially Weighted Moving Average: $\hat{\gamma}^{t+1} = (1-\alpha)\cdot\gamma^t + \alpha\cdot \mathrm{LN}(\mathrm{Concat}(h_1,\ldots,h_S)W^O)$

Layer normalization: $\gamma^{t+1} = \frac{\hat{\gamma}^{t+1}}{\sqrt{\sum_{j=1}^{M}(\hat{\gamma}_j^{t+1})^2}}$

Project $\gamma^{t+1} \in \mathbb{R}^{M\times D}$ into a dimension of E as attractors within associative memory: $f_{\mathrm{LT}}(\gamma^{t+1}) \in \mathbb{R}^{M\times E}$

Reconstruct Ξ using attractors $f_{\mathrm{LT}}(\gamma^{t+1})$ with a continuous Hopfield network in one inner step of energy reduction: $\hat{\Xi}^t = \arg\min_{\Xi^t}(-\mathrm{lse}(\beta, f_{\mathrm{LT}}(\gamma^{t+1})\Xi^t) + \frac{1}{2}\Xi^t{\Xi^t}^T + \beta^{-1}\log M + \frac{1}{2}(\max_i |f_{\mathrm{LT}}(\gamma_i^{t+1})|)^2)$

Reshape the tokens into batches as the output of the Global Workspace layer: $\hat{V}^t \in \mathbb{R}^{B\times N\times E} \leftarrow \hat{\Xi}^t \in \mathbb{R}^{(B\times N)\times E}$

Bottleneck Attention Balance Loss

Cumulative attention loss: $\ell_{\mathrm{importance}_{i,o}} = \sum_{j=1}^{M} \mathrm{A}_{i,j,o}$

Selected instance loss: $\ell_{\mathrm{loads}_{i,o}} = \sum_{j=1}^{M}(\mathrm{A}_{i,j,o} > 0)$

For each attention head i: $\ell_{\mathrm{bottleneck}_i} = \frac{\mathrm{Var}(\{\ell_{\mathrm{importance}_{i,o}}\}_{o=1}^{B\times N})}{(\frac{1}{B\times N}\sum_{o=1}^{B\times N}\ell_{\mathrm{importance}_{i,o}})^2+\epsilon} + \frac{\mathrm{Var}(\{\ell_{\mathrm{loads}_{i,o}}\}_{o=1}^{B\times N})}{(\frac{1}{B\times N}\sum_{o=1}^{B\times N}\ell_{\mathrm{loads}_{i,o}})^2+\epsilon}$

Sum the losses over all heads: $\sum_{i=1}^{S}\ell_{\mathrm{bottleneck}_i}$

Thank you!