Exploring Priors and Long-Term Memory in Transformers

Yuwei Sun

RIKEN AIP
Araya Research

Can Multimodal LLMs Learn by Memorizing More Data?

Probably not..

Q: Do you see any window in this image?

No, there are **no windows visible** in this image. The image shows a traditional red pillar box, which is a type of mailbox found in the United Kingdom.

Q: Is the door of the truck cab open?

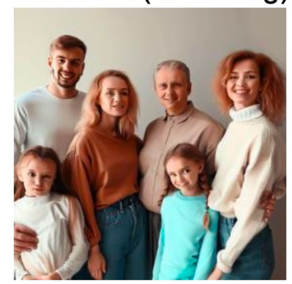
No, the door of the truck is **not open** in the image provided.

[Tong, 2024]

MidJourney 5.1

"an empty glass"

DALL-E (New Bing)



"a family of five members"

Stable Diffusion XL

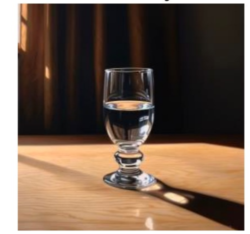
"a man descending a mountain"

Stable Diffusion 2.1

"there is no star in the night sky"

[Jones, 2023]

MidJourney 5.1



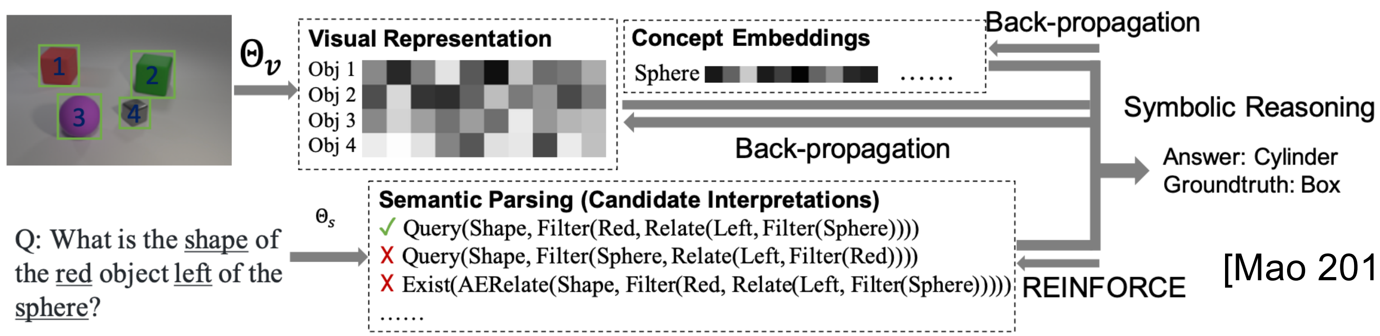
"an empty glass"

DALL-E (New Bing)

"a family of five members"

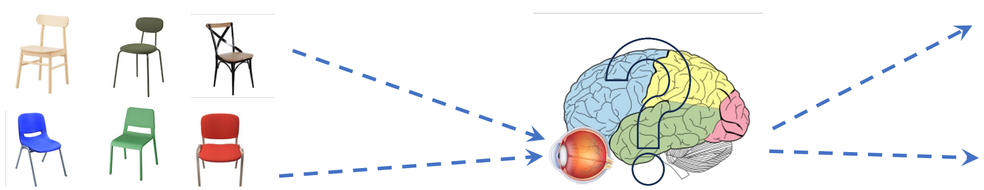
Neuro-symbolic is not very helpful for complex data

- ➢ Transformers are not good at learning discrete information from higher-dimensional perception data such as images, causing hallucination, inefficiency in training, and being data-hungry.
- ➢ Neuro-symbolic approach offers a more stable way to learn discrete symbols and their relations. However, the brain can learn without any annotated data, and real-world image data cannot be fully structured with a set of symbols.



[Mao 2019]

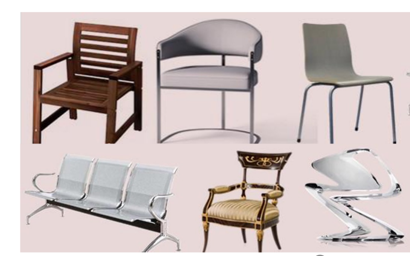
Unknown Mechanism of Induction



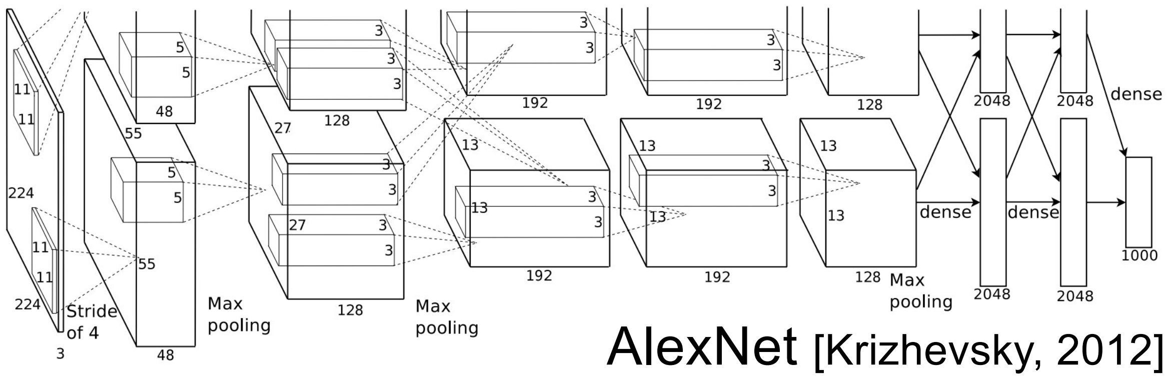
Color: black, brown,
Shape: square, round,
Material: wooden, ...
Azimuth: left, right, ...

Complexity of Image data

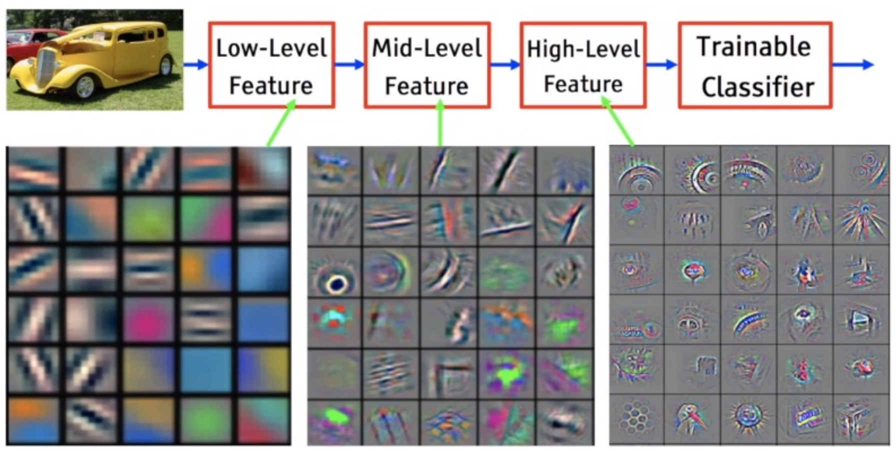
- Occlusion
- Complex Concept
- Complex Scene

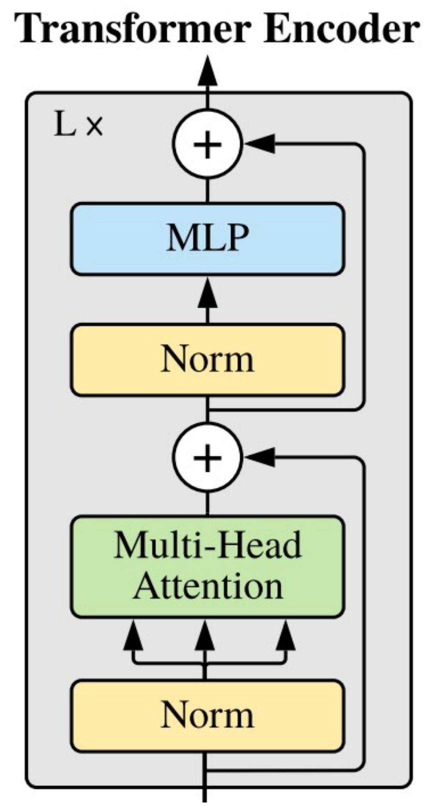


Absence of inductive biases such as convolution operations for localized knowledge in Transformers



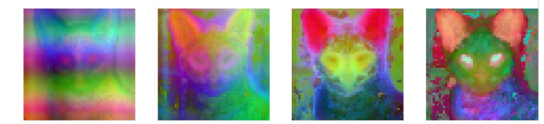
AlexNet [Krizhevsky, 2012]





[Dosovitskiy, 2021]

[Amir, 2022]



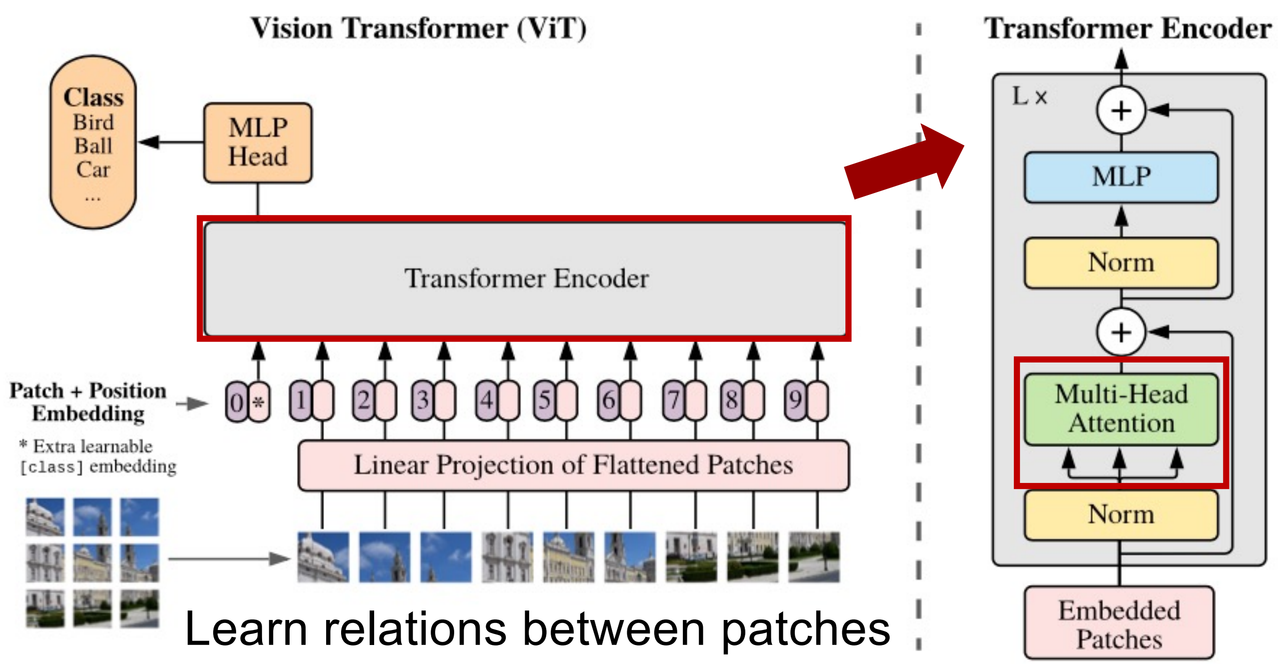
Low-level High-level

Do not learn structural features that align with the input data

➢ Difficult to generalize with limited samples

➢ Unlike convolution operations in CNNs, Transformers do not learn structural features that align with the input data and usually perform worse than CNNs with limited samples.

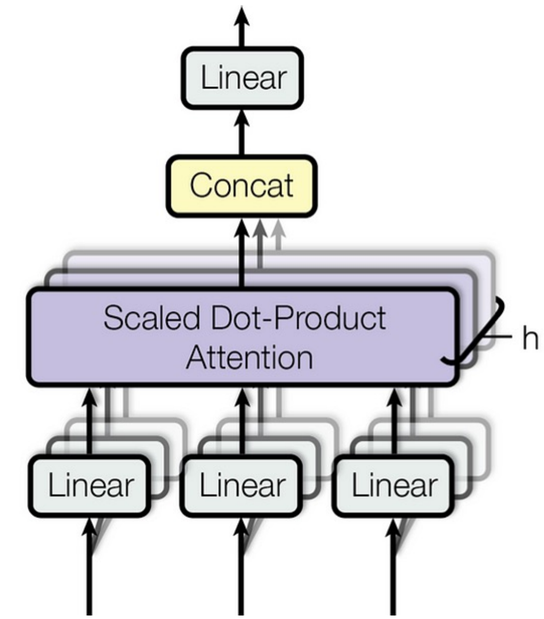
Vision Transformer and attention mechanism



Learn relations between patches

[Dosovitskiy, 2021]

Multi-Head Attention



V(Value) K(Key) Q(Query)

$$\mathbf{A_S} = \mathrm{softmax}\left(\frac{QW^{Q_i}(KW^{K_i})^T}{\sqrt{d}}\right),$$

$$\mathrm{Attention}(QW^{Q_i}, KW^{K_i}, VW^{V_i}) = \mathbf{A_S}\,\mathbf{V}$$

➢ Transformers use pairwise attention to establish correlations among disparate input segments.

*** Assign different weights to input tokens.**

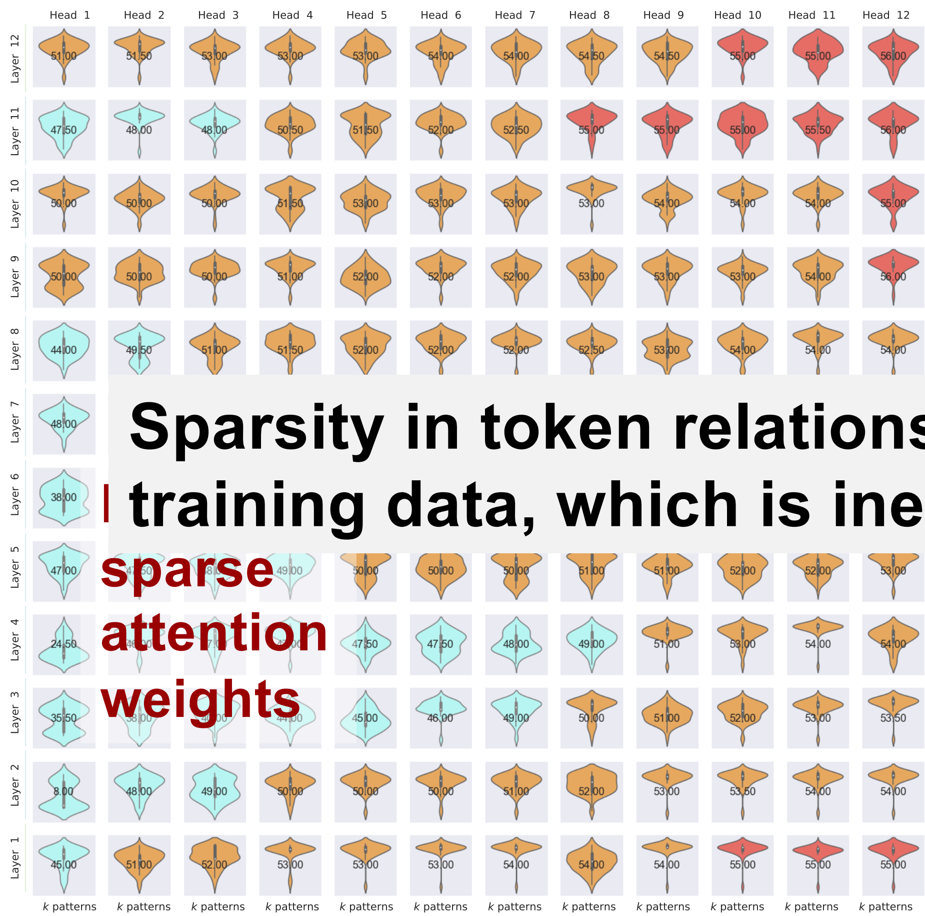
Analysis of attention weights in pretrained Transformers

Vision Transformer

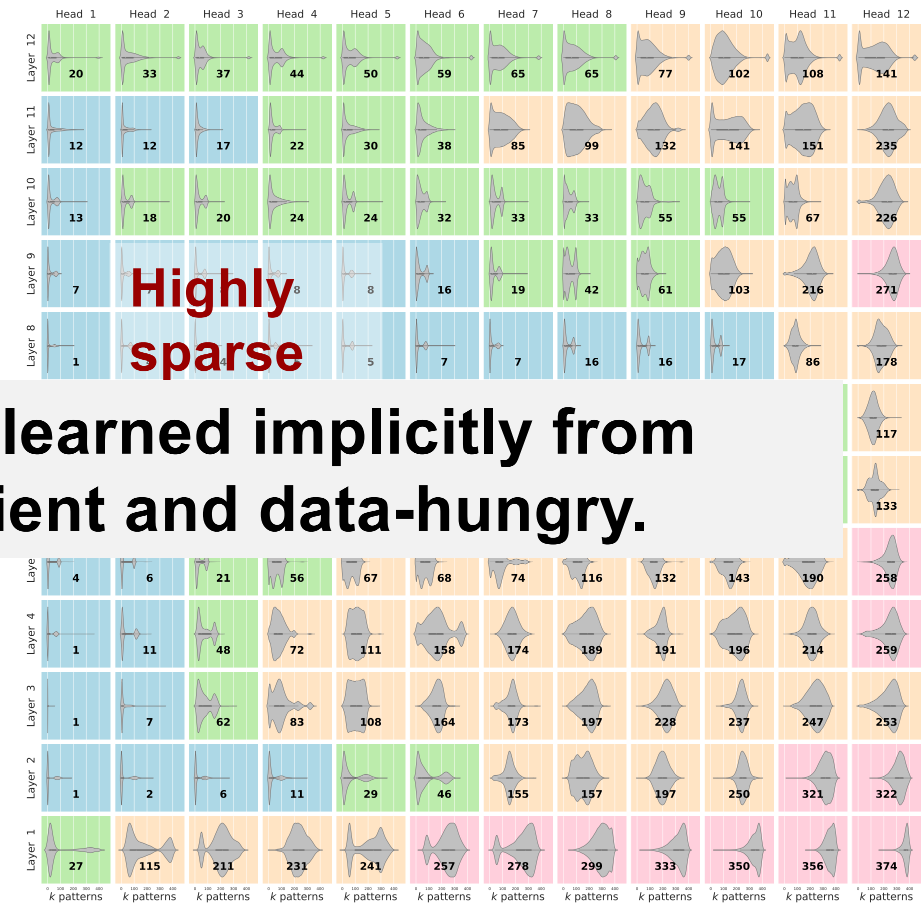
BERT (Natural language Transformer)

Layer 12

Layer 1



sparse attention weights



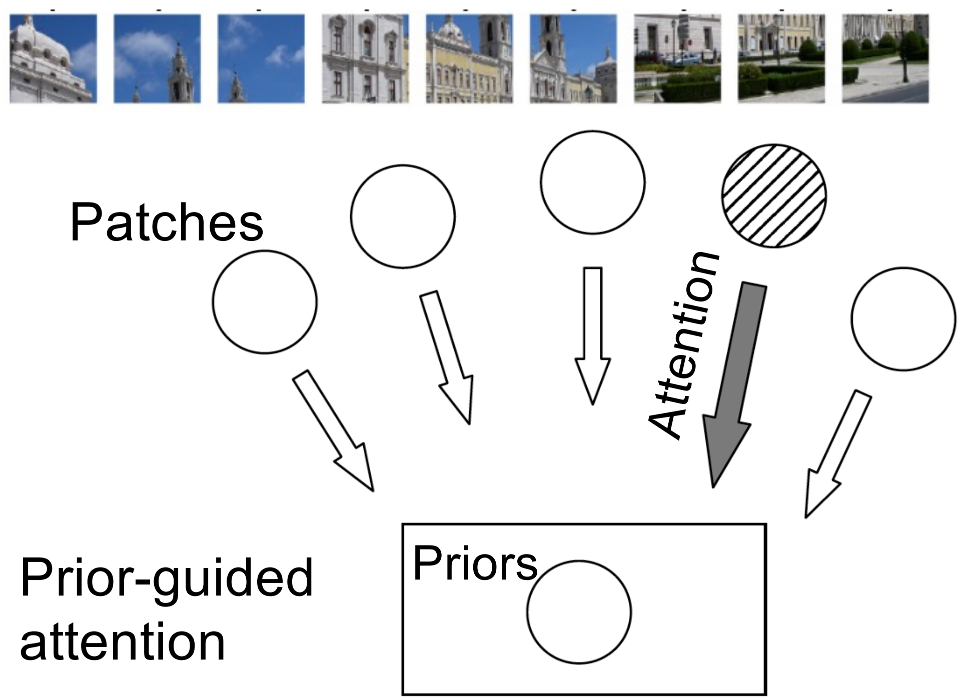
Highly sparse

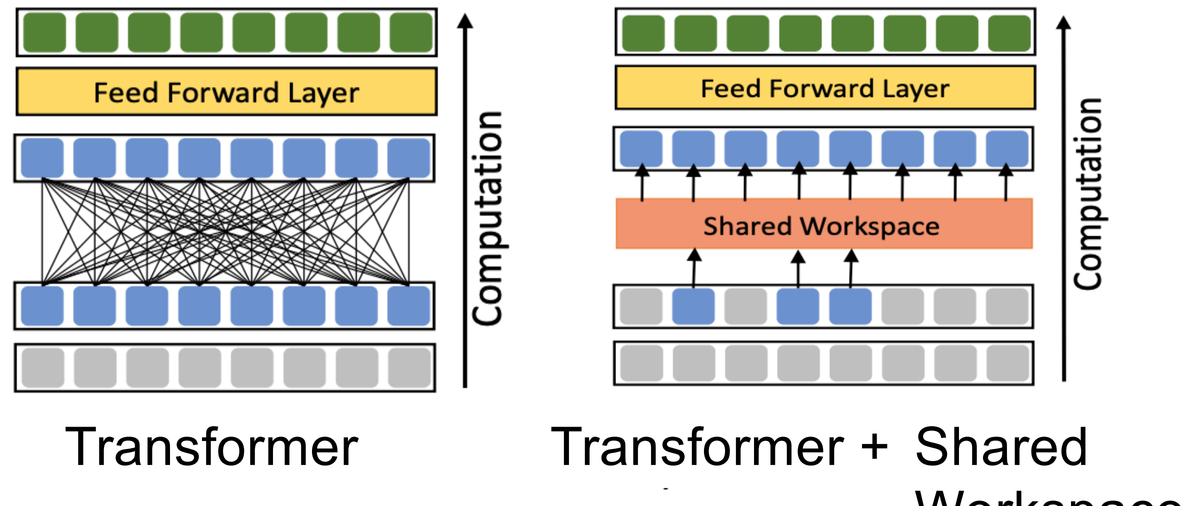
Sparsity in token relations is learned implicitly from training data, which is inefficient and data-hungry.

The attention sparsity is measured by the minimal number of required tokens whose attention scores add up to 0.90

[Ramsaue 2021; Sun 2023]

Inducing an information bottleneck in the attention mechanism





[Goyal et al. ICLR'22]

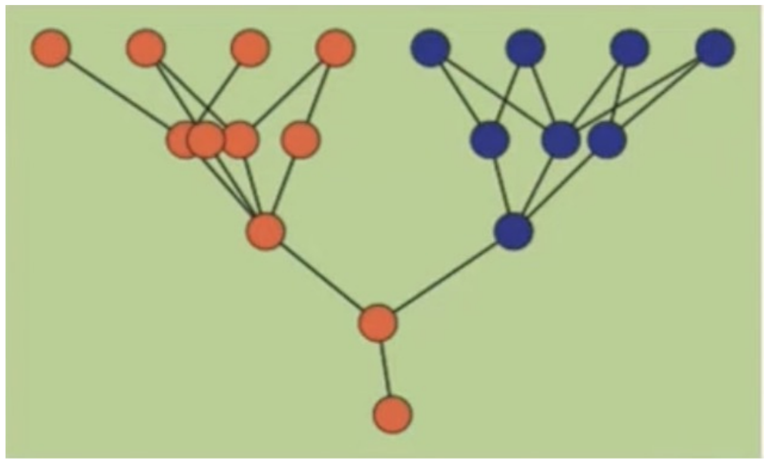
- Priors are general assumptions about samples, such as the aggregated features from different samples of the same object.
- Competition through a bottleneck results in naturally emerging specialized priors.

Specialized neural modules

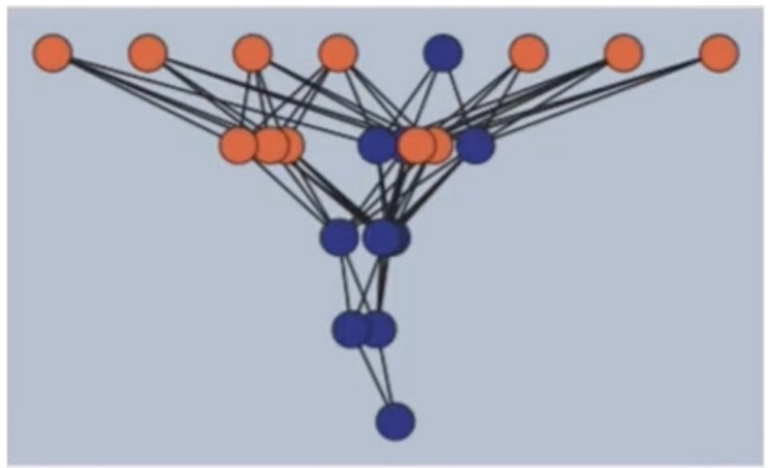
Definition: correspondence between strongly interconnected structural components of a network (modules) and the specialized functions they perform.

In animal brains, modularity favors evolvability, the ability to adapt to changing environments with common sub-problems [Clune, 2013]

Modular network



Non-modular network

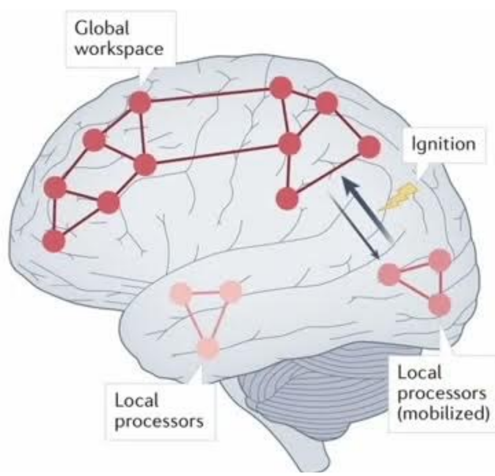


Working memory in the biological brain and Global Workspace Theory

When we use working memory: (1) Learning in a novel situation (2) Taking a different approach in familiar situations

Capacity: Limited amount of information it can hold at one time

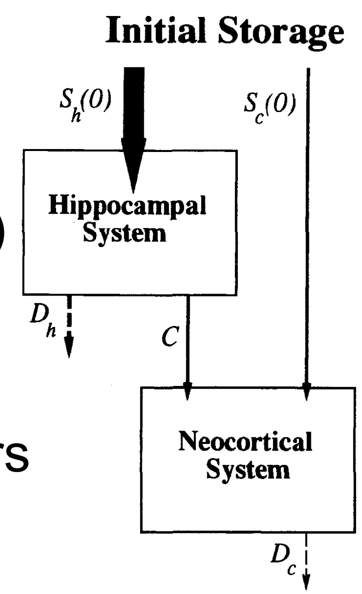
Function: Acts as a mental workspace where information can be manipulated: multiple specialized modules (potientially, multi-modal) compete to write to the shared space; information in the shared space is broadcast to all modules afterwards.



[Baars, 1988; Butlin, 2023]

Long-term memory
(Associative memory)

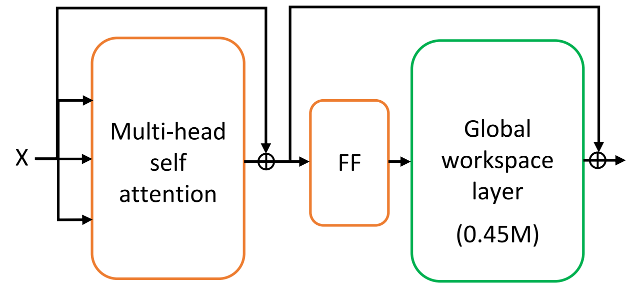
Reuse priors



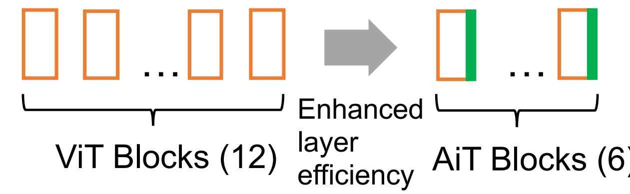
Learning priors

Working memory
and Long-term memory
(Semantic memory)

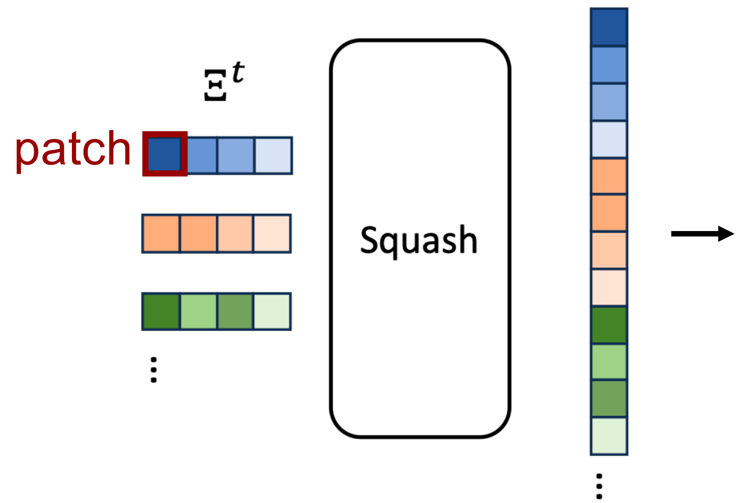
Global Workspace Layer



Smaller computational cost

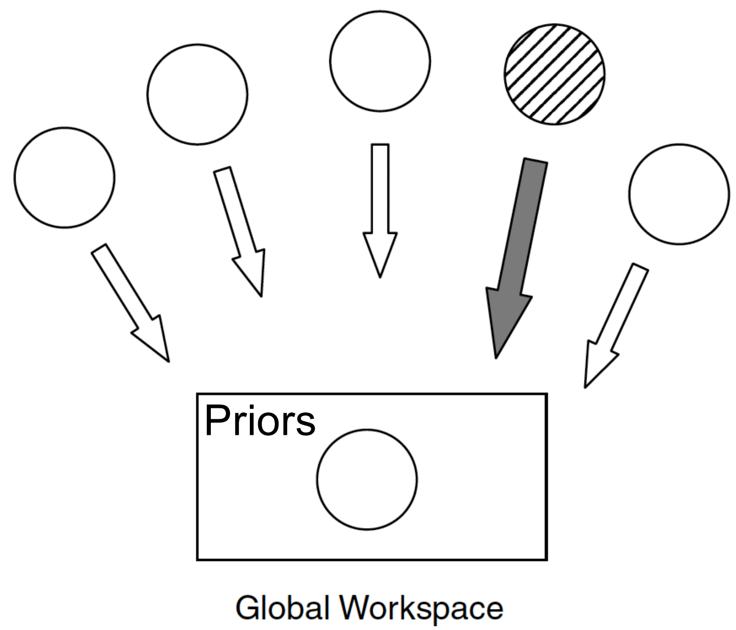


(1) Collecting patches from all batch samples

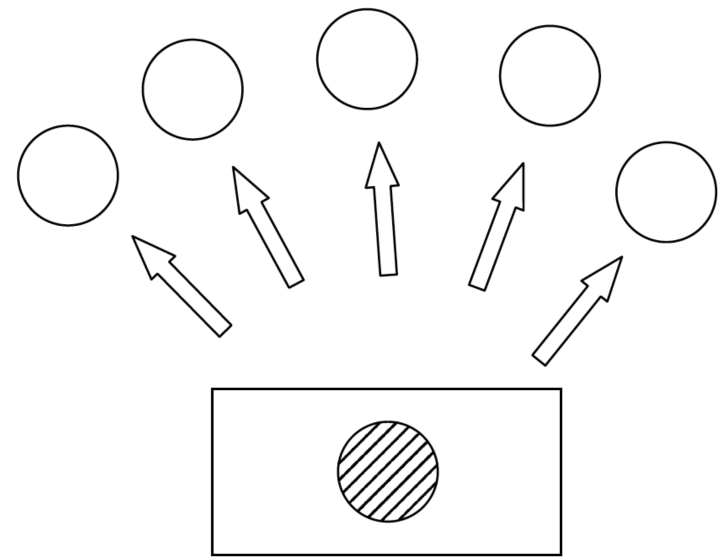


Images are represented by different hues. Patches from the same image are distinguished by differences in brightness.

(2) Computing the bottleneck attention and selecting relevant patches

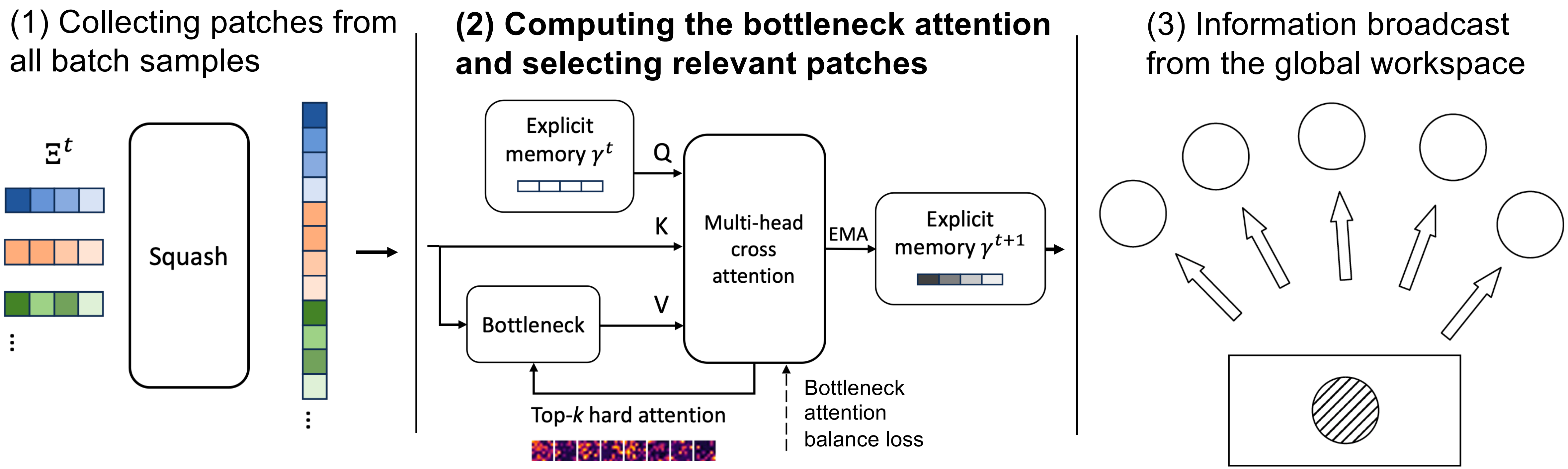


(3) Information broadcast from the global workspace



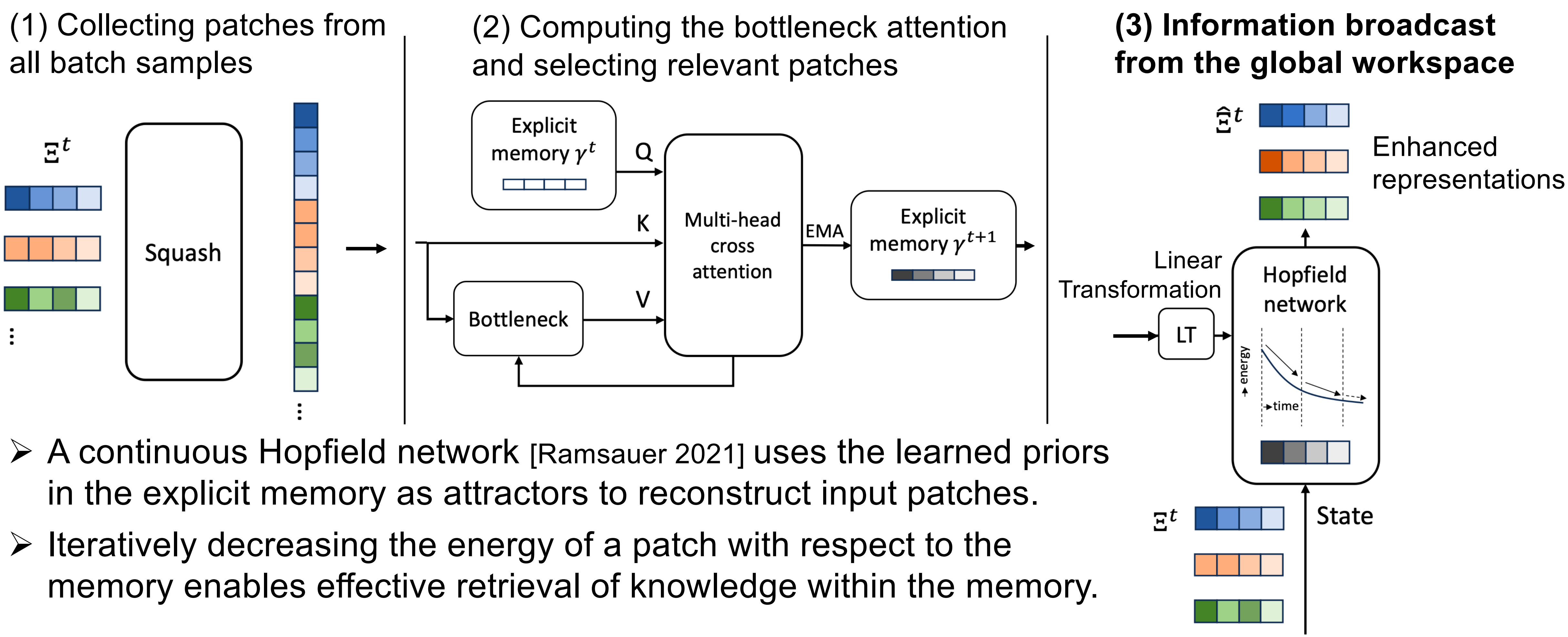
Sun et al., Associative Transformer, NeurIPS workshop; arXiv:2309.12862

Global Workspace Layer



- The explicit memory stores and updates a set of priors (as queries) by attending to different patches based on the multi-head cross attention.
- The sparsity is enabled through a bottleneck using the top-*k* hard attention.

Global Workspace Layer

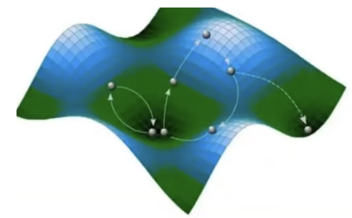


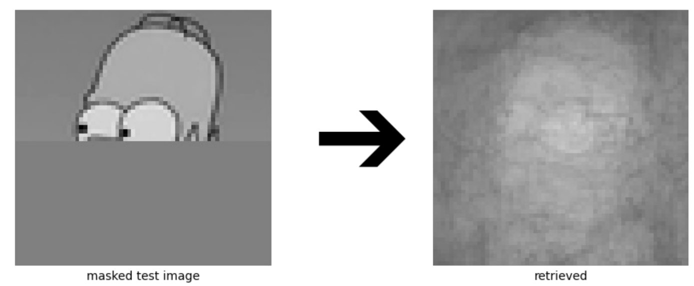
➢ A continuous Hopfield network [Ramsauer 2021] uses the learned priors in the explicit memory as attractors to reconstruct input patches.

➢ Iteratively decreasing the energy of a patch with respect to the memory enables effective retrieval of knowledge within the memory.

Continuous Hopfield Network

[Ramsauer, 2021]





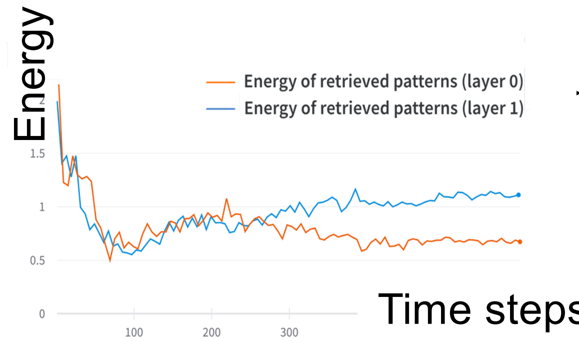
Multiple interactions of energy reduction to reconstruct patterns

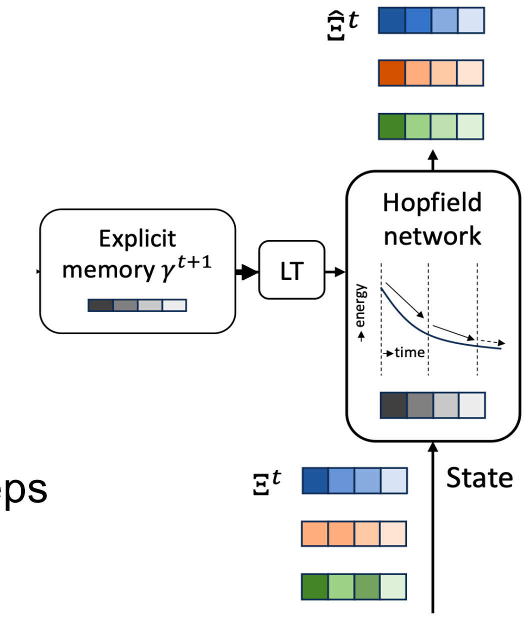
Token retrieval with a continuous Hopfield network

- Energy of the continuous Hopfield network

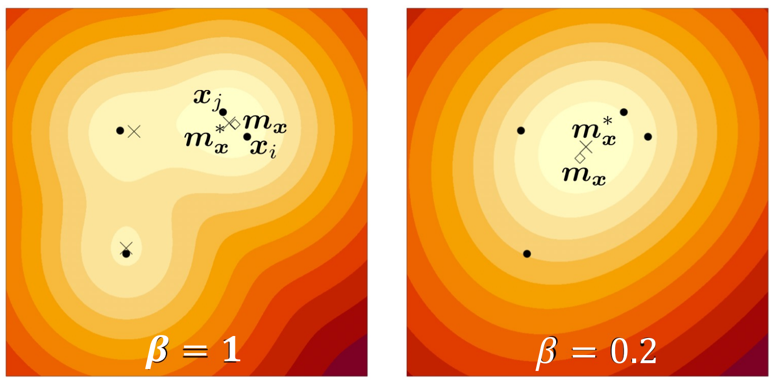
$$E(\xi^t) = -\mathrm{lse}(\boxed{\beta}, f_{\mathrm{LT}}(\gamma^{t+1})\xi^t) + \frac{1}{2}\xi^t {\xi^t}^T + \boxed{\beta}^{-1}\log M + \frac{1}{2}\zeta^2$$

$$\zeta = \max_i |f_{\mathrm{LT}}(\gamma_i^{t+1})|, \quad \hat{\xi}^t = \arg\min_{\xi^t} E(\xi^t)$$





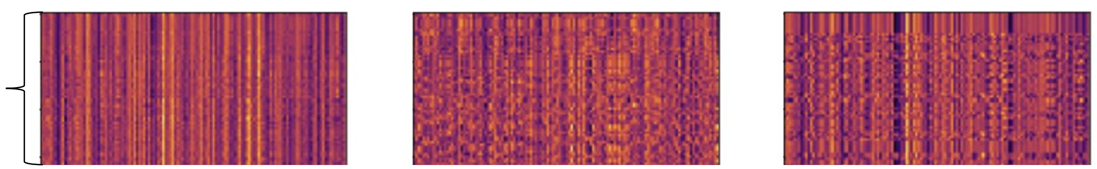
- Inverse temperature β and basins of attraction



[Ramsauer, 2021]

Reconstructed features from memory

Identical patches retrieved



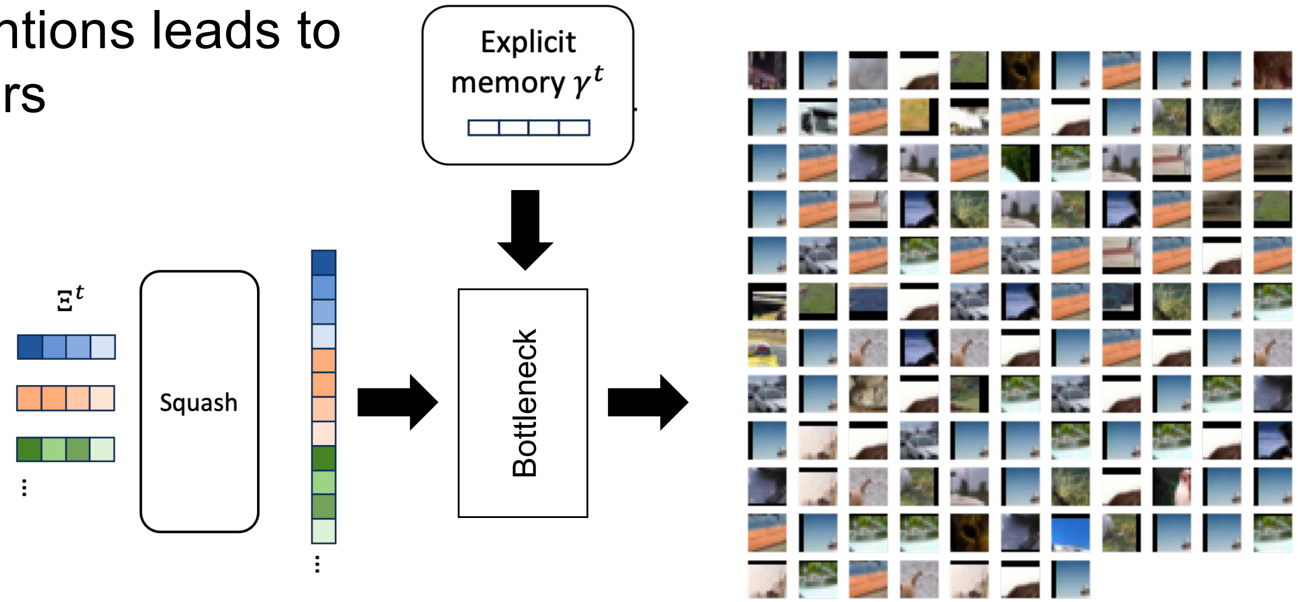
β = 0.05 β = 1.0 β = 4.0

- A smaller β results in a metastable state within the basin of multiple attractors.

- A very large or small β both can lead to local minima.

Problem 1: monolithic priors select repeated tokens: Introducing Bottleneck Attention Balance Loss

- Cascading multiple bottleneck attentions leads to difficulty in forming specialized priors



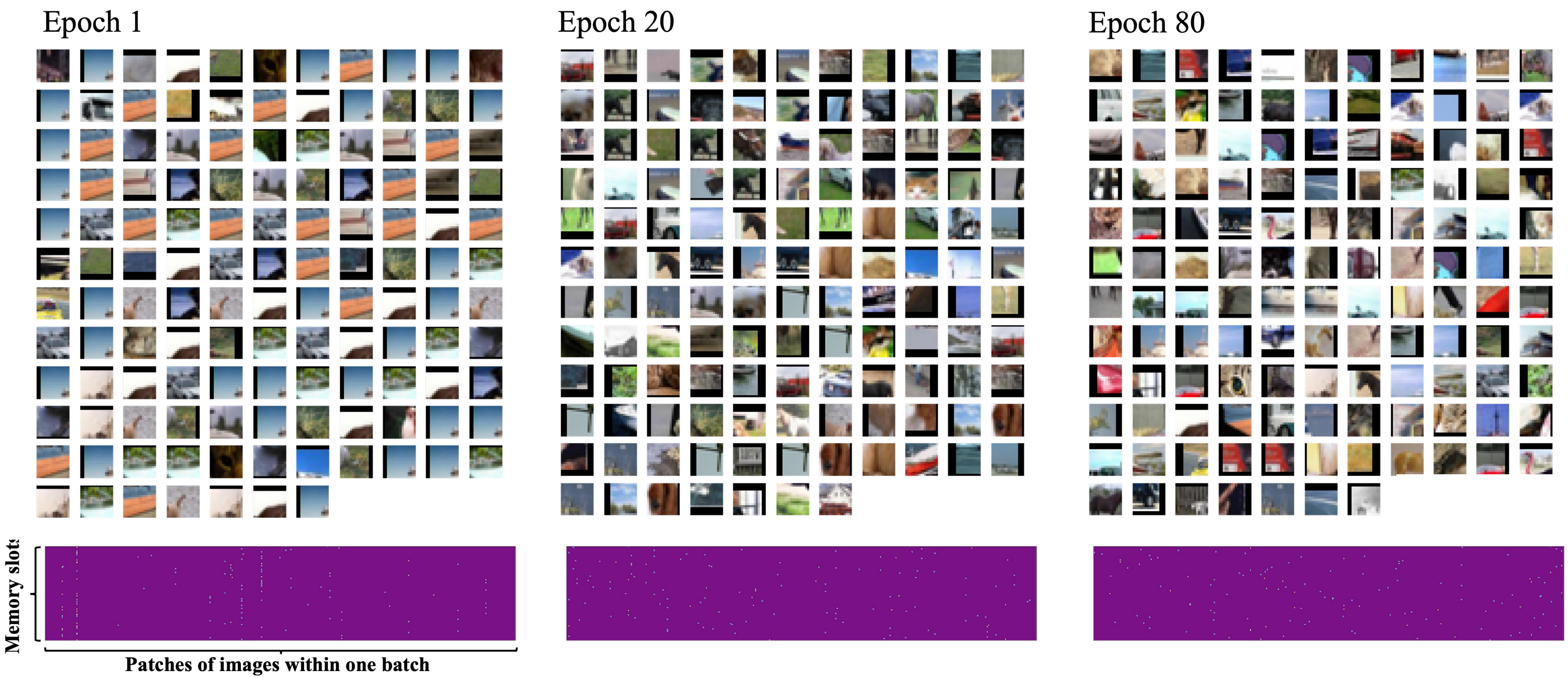
Cumulative attention loss: $\ell_{\text{importance}_{i,o}} = \sum_{j=1}^{M} \mathrm{A}_{i,j,o}$

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

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

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

Diversity in patch selection with the new loss



Problem 2: Computational load with the squash operation

The squash layer concatenates all tokens in the batch, $\Xi \in \mathbb{R}^{(B\times N)\times E}$, allowing for across-sample learning but also increasing the computational cost for the attention mechanism.

To reduce its the computational load:
(1) a low rank memory, where the squashed representations are projected to a latent space of dimension $D \ll E$

(2) an attention bottleneck with capacity $k \ll B\times N, \mathrm{e.g.},$ 1.6% ~ 3.2% of all the tokens in our experiments

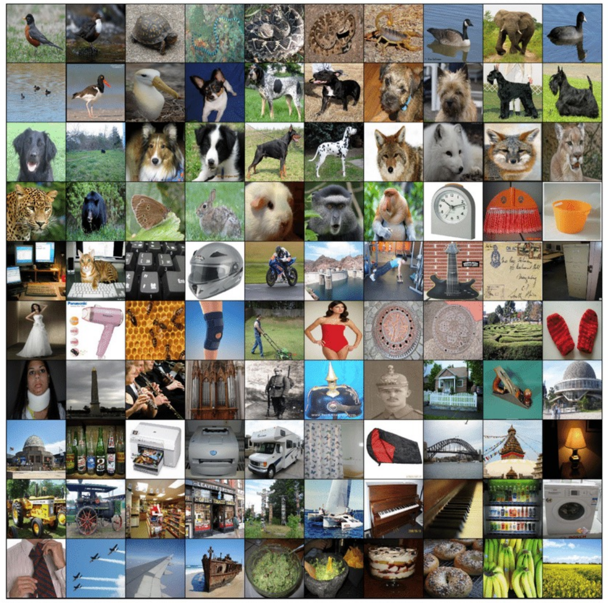
Methods	Size (M)	#FLOPs
AiT-Base	91.0	5.77×10^9
AiT-Small	15.8	9.64×10^8
ViT-Base	85.7	5.60×10^9
ViT-Smal	14.9	9.36×10^8

- Less than a 3% increase in computation compared to Vision Transformers of similar size.

Enhanced efficiency in image classification tasks

Methods	CIFAR10	CIFAR100	Triangle	Average	Parameters (M)
AiT-Base	**85.44**	**60.78**	99.59	**81.94**	91.0
AiT-Medium	84.59	60.58	99.57	81.58	45.9
AiT-Small 6 layers	83.34	56.30	99.47	79.70	15.8
Coordination Goyal et al. (2022b)	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 Dosovitskiy et al. (2021)	83.82	57.92	**99.63**	80.46	85.7
ViT-Small 12 layers	79.53	53.19	99.47	77.40	14.9
Perceiver Jaegle et al. (2021)	82.52	52.64	96.78	77.31	44.9
Set Transformer Lee et al. (2019)	73.42	40.19	60.31	57.97	2.2
BRIMs Mittal et al. (2020)	60.10	31.75	-	45.93	4.4
Luna Ma et al. (2021)	47.86	23.38	-	35.62	77.6

ImageNet100

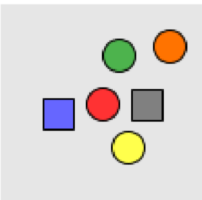


Methods	Test accuracy (%)	Size (M)
AiT-Medium	**36.72**	45.9
AiT-Small	33.84	15.8
ViT-Base	34.62	85.7
ViT-Medium	31.72	42.7
ViT-Small	28.16	14.9

Our study demonstrates that AiT outperforms existing sparse Transformer models including the variants of Coordination [Goyal 2022] and Vision Transformers, without pretraining on external data.

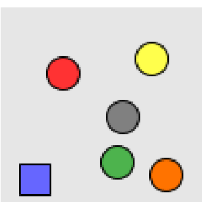
Vision-language relational reasoning tasks

Sort-of-CLEVR dataset [Santoro, 2017]



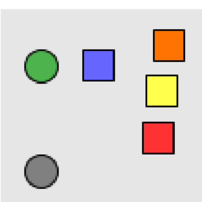
Non-relational question
Q: Is the yellow object on the top or on the bottom?
A: bottom

Relational question
Q: What is the color of the object that is closest to the blue object?
A: red



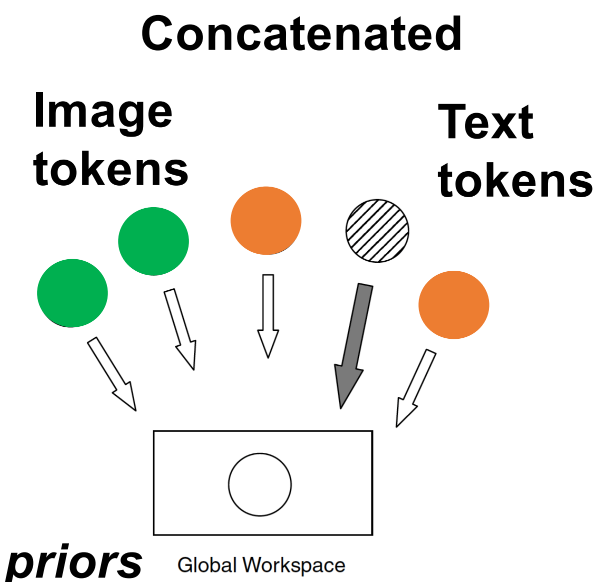
Non-relational question
Q: What is the shape of the red object?
A: circle

Relational question
Q: How many objects have the shape of the blue object?
A: 1



Non-relational question
Q: Is the blue object on the top or on the bottom?
A: top

Relational question
Q: What is the color of the object that is closest to the red object?
A: yellow



Cross-modal priors

Methods	Relational	Non-relational
Transformer based models		
AiT-Base	**80.03**	**99.98**
AiT-Medium	78.14	99.75
AiT-Small	76.82	99.85
Coordination	73.43	96.31
ViT-Base	63.35	99.73
ViT-Medium	54.71	99.70
ViT-Small	51.75	98.80
Set Transformer	47.63	57.65

Conclusions

- **Associative Transformer enhances parameter efficiency in the training of Transformer-based models, making them more accessible and cost-effective**
- **Implementing the cognitive science theory of the Global Workspace is crucial for a better understanding of human-like relational reasoning**
- Other tasks and domains, such as audio and video
- Safe deployment in real-world applications.

Privacy of neural module learning	Adversarial attacks
Bidirectional Contrastive Split Learning Sun et al. AAAI 2024	Attacking Distance-aware Attack Sun et al. Transactions on AI 2023

Exploring Priors and Long-Term Memory in Transformers

Yuwei Sun

RIKEN AIP
Araya Research

Associative Transformer paper

Center for Advanced Intelligence Project
ARAYA
Microsoft Research 微软亚洲研究院