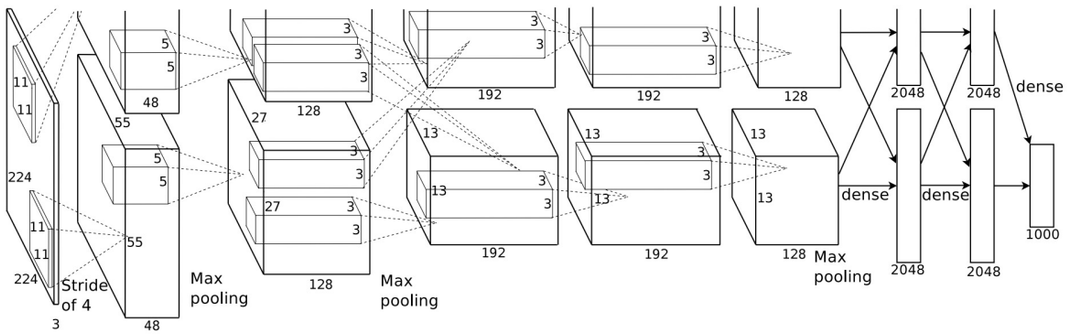
Associative Transformer Is A Sparse Representation Learner

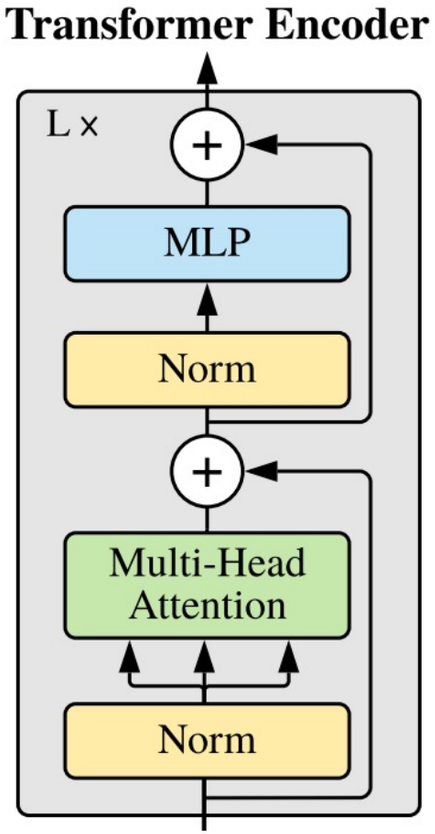
Yuwei Sun, Hideya Ochiai, Zhirong Wu, Stephen Lin, Ryota Kanai

Inductive bias allowing a model to generalize more effectively to unseen samples

AlexNet [Krizhevsky, 2012]

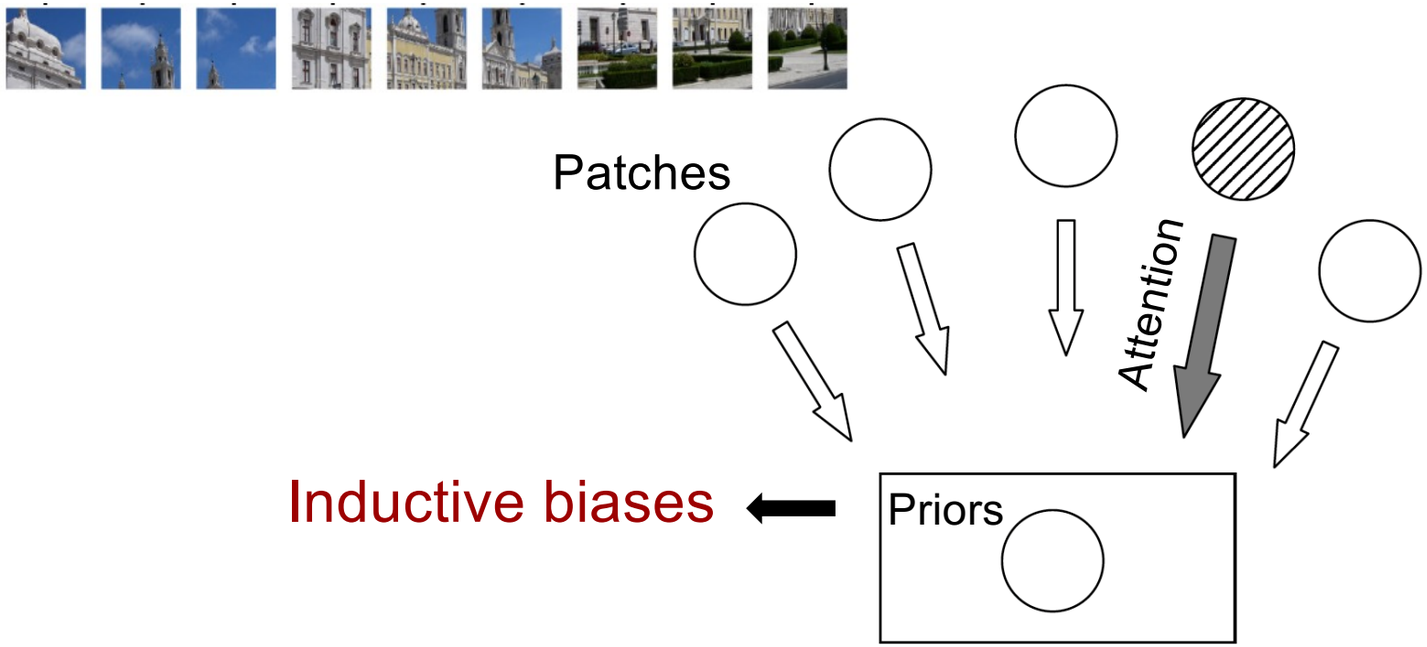


- Transformers use pairwise attention to establish correlations among disparate input segments
- Unlike convolution operations in CNNs, self-attention in Transformers does not possess an inductive bias that is consistent with the underlying input data structure



[Dosovitskiy, 2021]

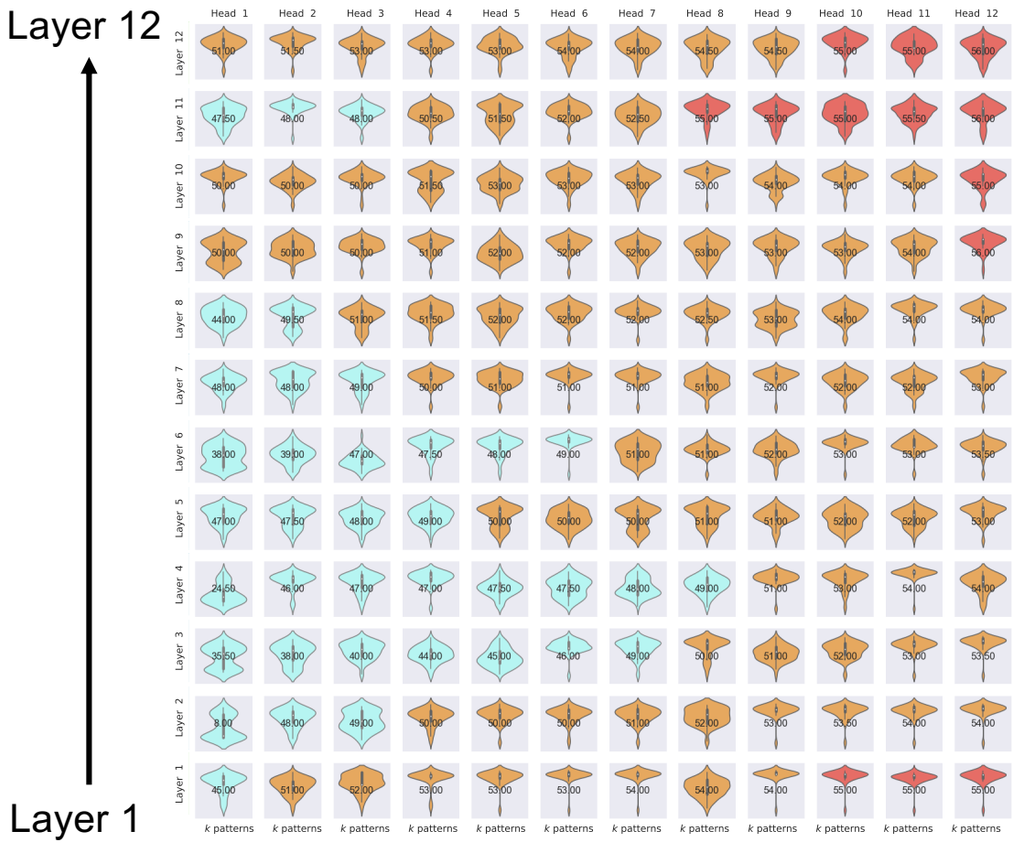
Emergent inductive biases through competition



- Absence of a bottleneck for localized, and contextual representation learning
- Competition results in naturally emerging specialized priors
- Self-attention in Transformers becomes expensive with scaling

Does competition exist within the self-attention mechanism?

Attention heads in Vision Transformer (Ours)



Attention heads in BERT [Ramsaue, 2021]

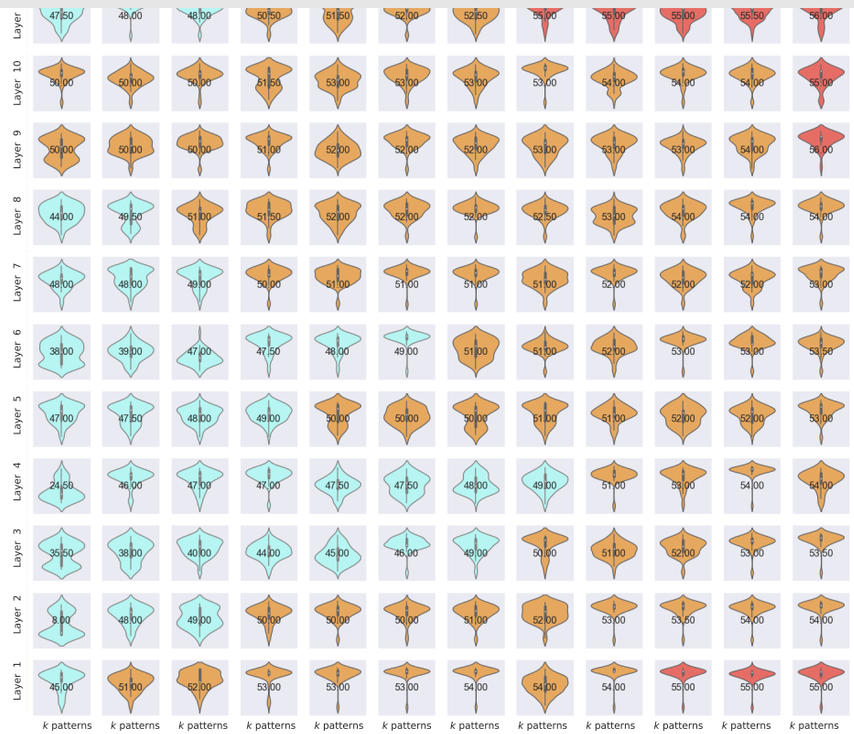
Does competition exist within the self-attention mechanism?

Competition, which reveals naturally sparser interactions among attention heads in pairwise attention, is important for learning meaningful representations.

Learning a set of priors that can guide attention in the learning process.

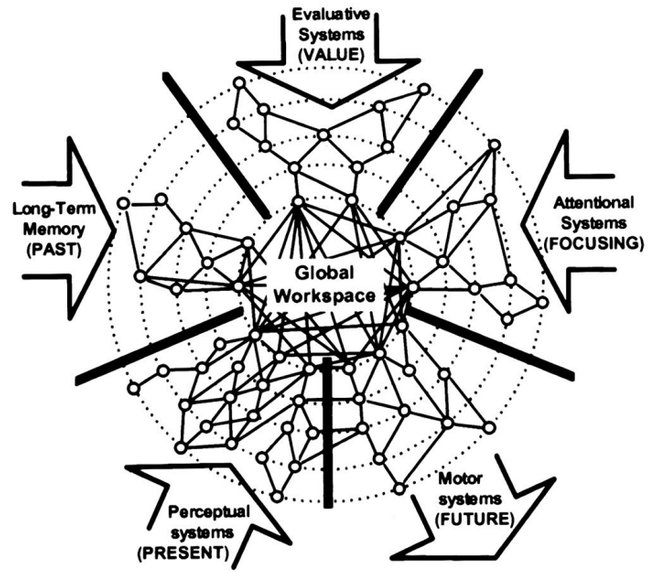
Lay

Layer 1



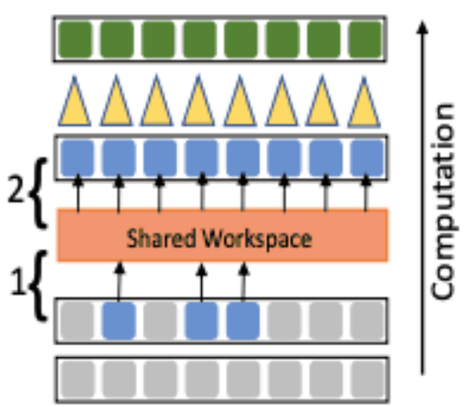
Inspired by the Global Workspace Theory (GWT)

- Guided attention in a limited capacity workspace
- Modules compete to write into the shared space
- Availability of information in the workspace through broadcast

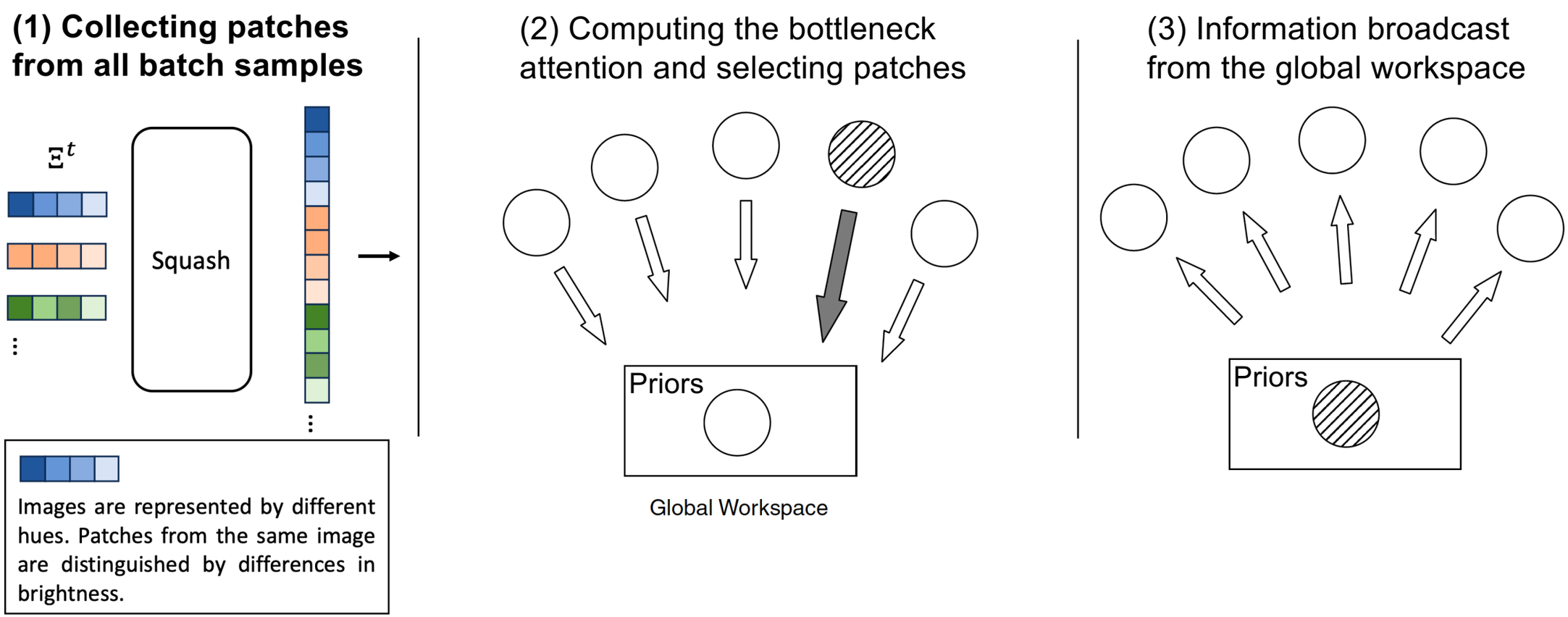


[Baars, 1988; Dehaene, 1998]

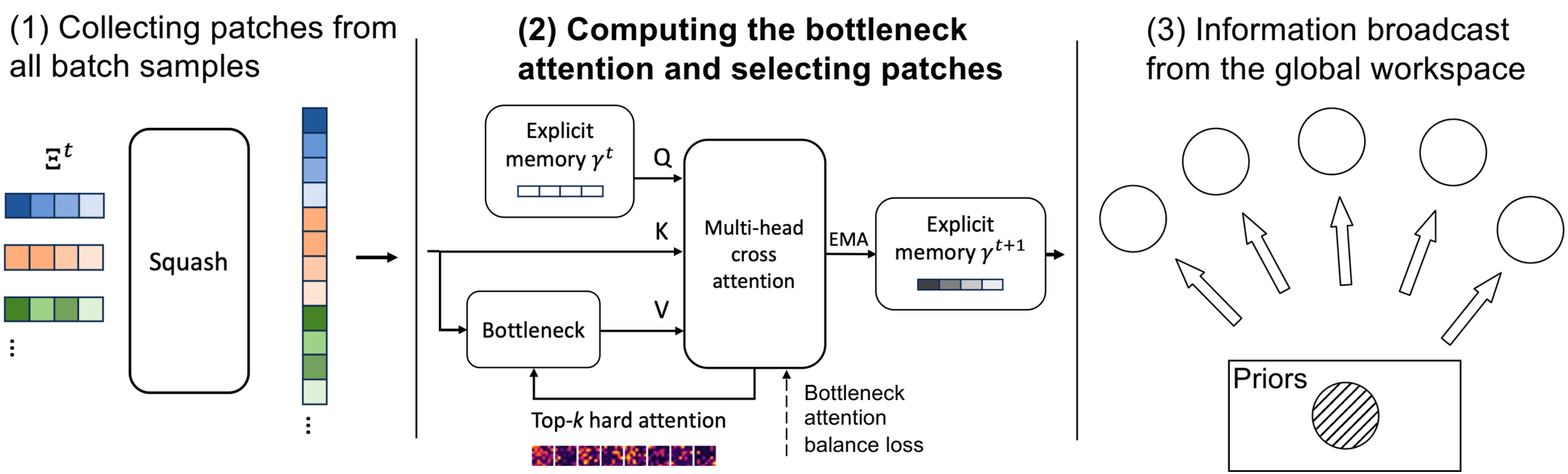
➢ The Coordination method functions as a Global Workspace based on cross-attention over input segments of a sample
[Goyal, ICLR 2022]



Inductive biases through prior specialization in global workspace

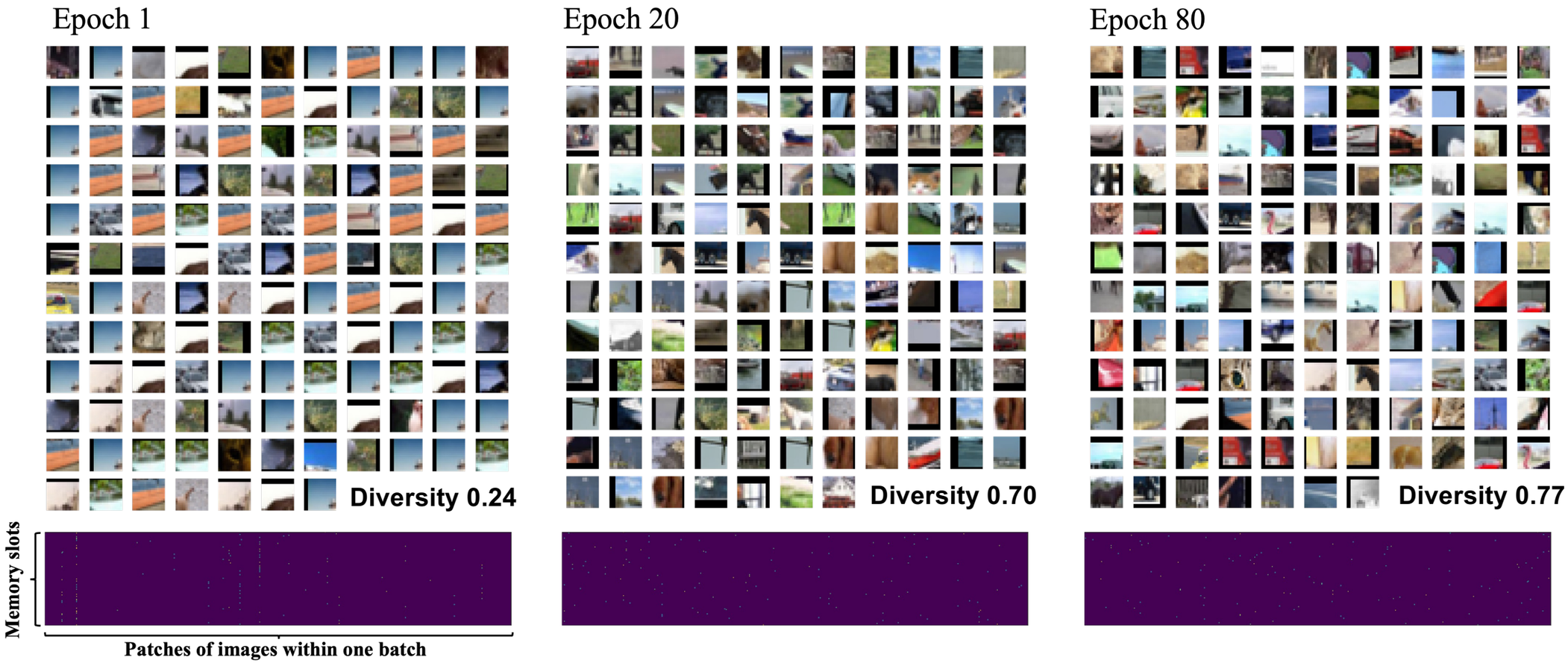


Inductive biases through prior specialization in global workspace

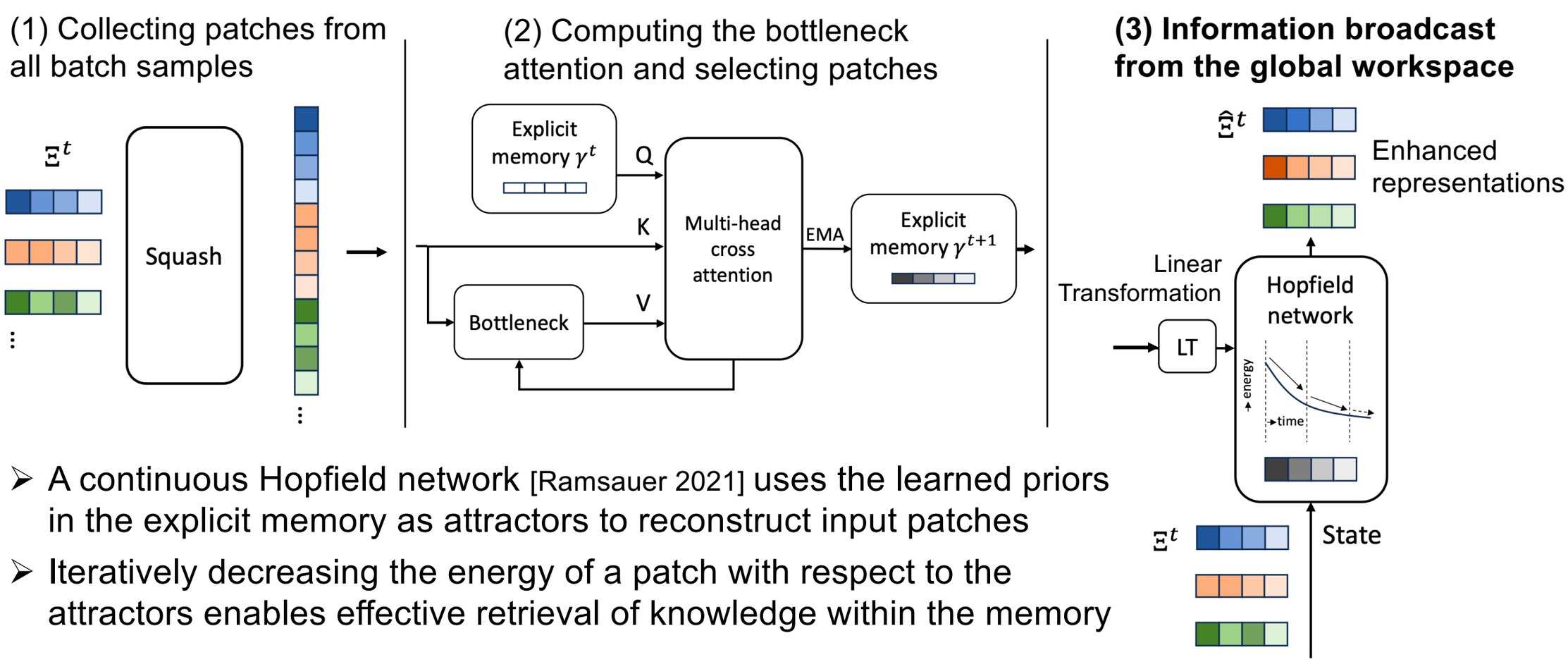


- The explicit memory stores and updates a set of priors by attending to different patches, which are smoothed by Exponential Moving Average (EMA)
- The balance loss encourages diverse patch selection of the Top-*k* hard attention

Selected patches by the specialized priors



Inductive biases through prior specialization in global workspace



- 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 attractors enables effective retrieval of knowledge within the memory

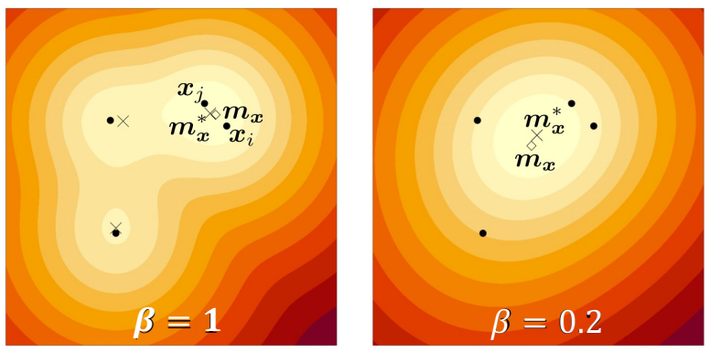
Associative memory in a continuous Hopfield network

- Energy of the continuous Hopfield network

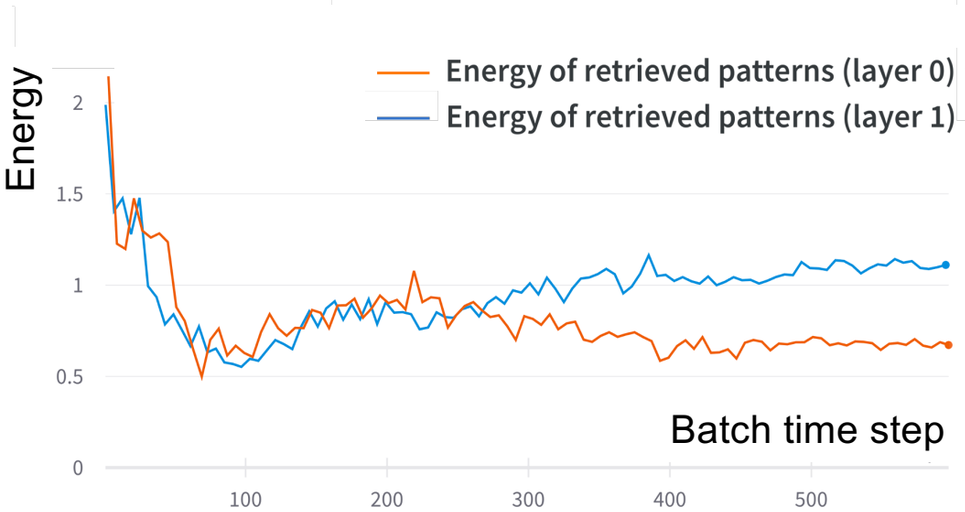
$$E(\xi^t) = -\mathrm{lse}(\beta, f_{\mathrm{LT}}(\gamma^{t+1})\xi^t) + \frac{1}{2}\xi^t\xi^{t^T} + \beta^{-1}\log M + \frac{1}{2}\zeta^2$$

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

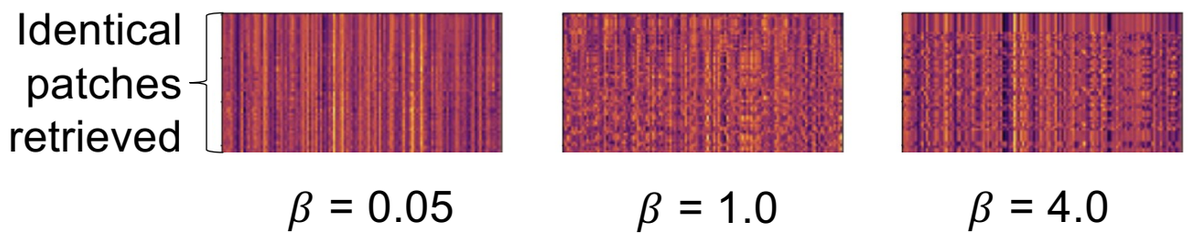
Inverse temperature β and basins of attraction



[Ramsauer, 2021]

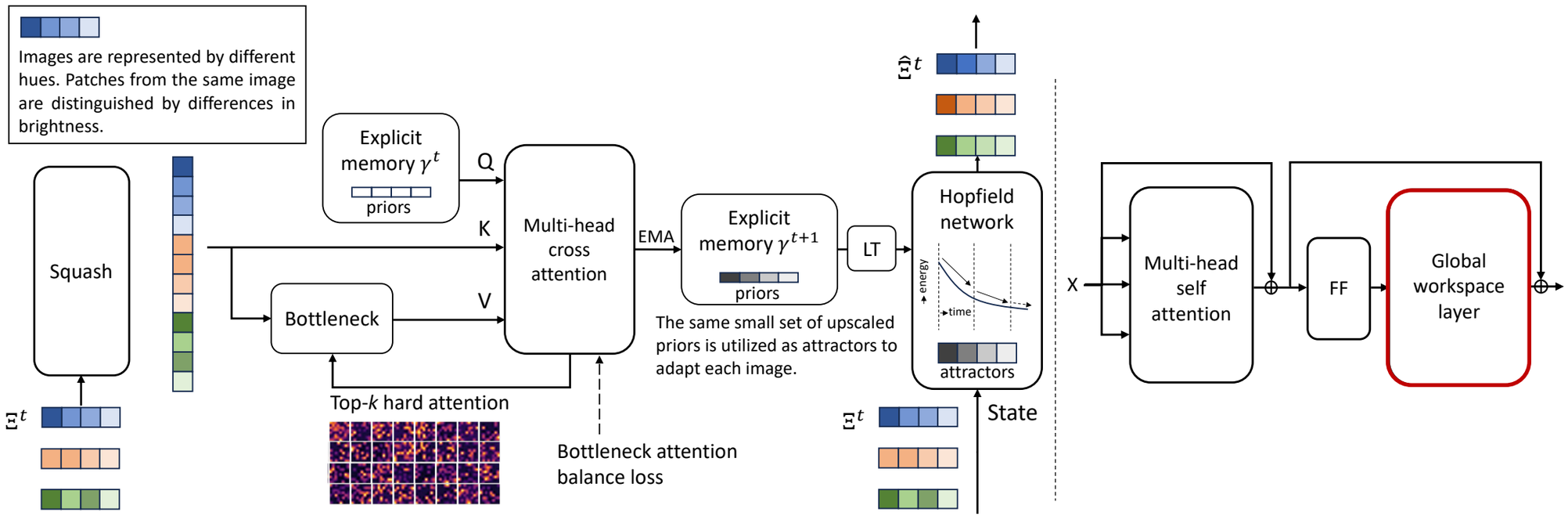


Reconstructed sample representations



- A smaller β is more likely to result in a metastable state

Associative Transformer

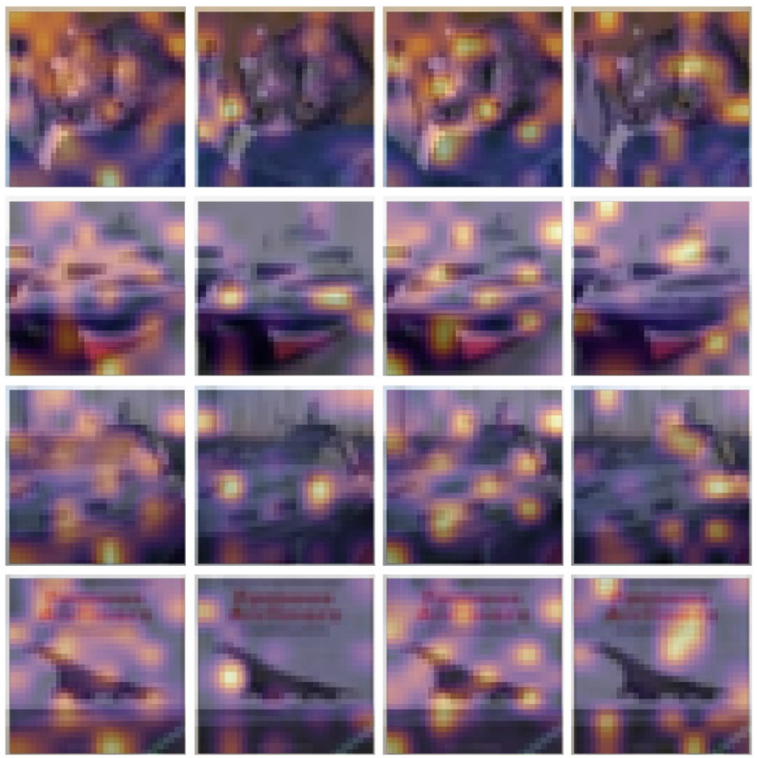


(a) Global workspace layer

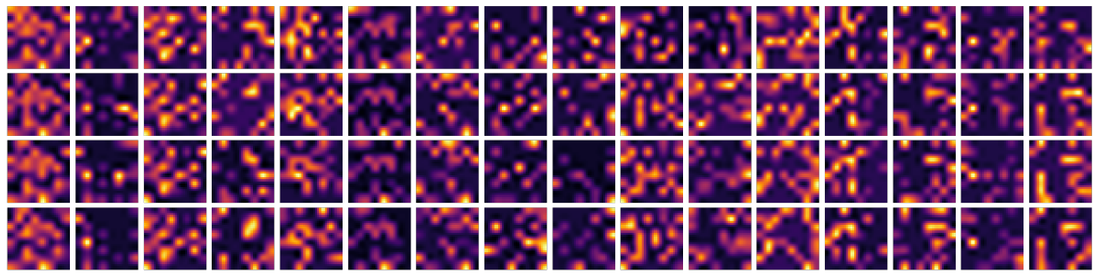
(b) Associative Transformer block

Prior specialization

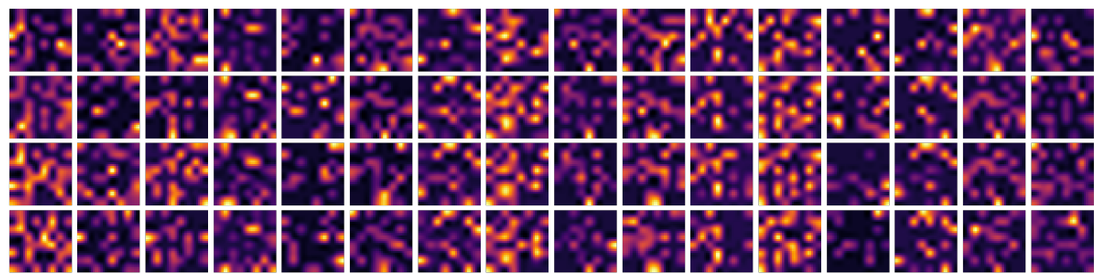
- Patches across samples are attended sparsely by different priors, with emergent specialization in a prior's attention



Prior 1 Prior 2 Prior 3 Prior 4



Prior 1-16



Prior 17-32

AiT outperforms latent memory-based Transformers

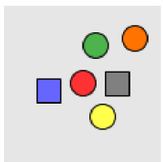
Methods	CIFAR10	CIFAR100	Triangle	Average	Model Size
AiT-Base	**85.44**	**60.78**	99.59	**81.94**	91.0
AiT-Medium	84.59	60.58	99.57	81.58	45.9
AiT-Small	83.34	56.30	99.47	79.70	15.8
Coordination Goyal et al. (2022b)	75.31	43.90	91.66	70.29	2.2
Coordination-DH	72.49	51.70	81.78	68.66	16.6
Coordination-D	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	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

AiT is a layer-efficient model

	Methods	CIFAR10	CIFAR100	Triangle	Average	Model Size
	AiT-Base	**85.44**	**60.78**	99.59	**81.94**	91.0
6 layers	AiT-Medium	84.59	60.58	99.57	81.58	45.9
	AiT-Small	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
12 layers	ViT-Base Dosovitskiy et al. (2021)	83.82	57.92	**99.63**	80.46	85.7
	ViT-Small	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

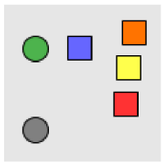
Multi-modal relational reasoning tasks

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

Methods	Relational	Non-relational	Average
Transformer based models			
AiT-Small	**76.82**	**99.85**	**88.34**
Coordination Goyal et al. (2022b)	73.43	96.31	84.87
Set Transformer Lee et al. (2019)	47.63	57.65	52.64
Non-Transformer based models (upper bound)			
CNN+RN Santoro et al. (2017)	81.07	98.82	89.95
CNN+MLP Santoro et al. (2017)	60.08	99.47	79.78
Dense-Base	46.93	57.71	52.32
Dense-Small	47.28	57.68	52.49

- AiT achieved competitive performance with the upper bound non-Transformer based models that rely on a built-in inductive bias, through the efficient association of disparate information fragments.

Conclusions

- Associative Transformer (AiT): a biologically plausible sparse attention mechanism based on the Global Workspace Theory and associative memory
- AiT is a sparse representation learner, leveraging bottleneck attention to acquire specialized priors
- Adaptive low-rank priors increase memory capacity, allowing AiT to learn up to 128 specialized priors from a diverse pool of 32.8k patches (refer to the paper)
- The learned priors serve as attractors in a Hopfield network: The first work to incorporate the Hopfield network as an integral element in a sparse attention mechanism for inductive biases

NEURAL INFORMATION PROCESSING SYSTEMS

Associative Transformer Is A Sparse Representation Learner

Yuwei Sun, Hideya Ochiai, Zhirong Wu, Stephen Lin, Ryota Kanai

東京大学 THE UNIVERSITY OF TOKYO

ARAYA

Microsoft Research 微软亚洲研究院