Localized Learning Through the Lens of Global Workspace Theory

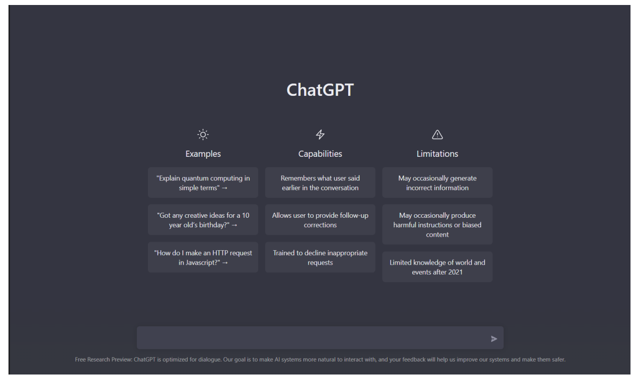
Yuwei Sun

University of Tokyo

The state of deep learning

Go

Large language model



Self-driving car

Text to video

Protein folding

Building systems with consciousness?

8v2 [cs.AI] 21 Aug 2023

Consciousness in Artificial Intelligence:
Insights from the Science of Consciousness

Patrick Butlin* **Robert Long*** **Eric Elmoznino**
Yoshua Bengio **Jonathan Birch** **Axel Constant**
George Deane **Stephen M. Fleming** **Chris Frith**
Xu Ji **Ryota Kanai** **Colin Klein**
Grace Lindsay **Matthias Michel** **Liad Mudrik**
Megan A. K. Peters **Eric Schwitzgebel** **Jonathan Simon**
Rufin VanRullen

Abstract

Whether current or near-term AI systems could be conscious is a topic of scientific interest and increasing public concern. This report argues for, and exemplifies, a rigorous and empirically grounded approach to AI consciousness: assessing existing AI systems in detail, in light of our best-supported neuroscientific theories of consciousness. We survey several prominent scientific theories of consciousness, including recurrent processing theory, global workspace theory, higher-order theories, predictive processing, and attention schema theory. From these theories we derive ”indicator properties” of consciousness, elucidated in computational terms that allow us to assess AI systems for these properties. We use these indicator properties to assess several recent AI systems, and we discuss how future systems might implement them. Our analysis suggests that no current AI systems are conscious, but also shows that there are no obvious barriers to building conscious AI systems.

Making sense of information processing

- **Selection of information for global broadcasting, thus making it flexibly available for computation and report (C1)**

- Self-monitoring of those computations, leading to a subjective sense of certainty or error (C2)

Dehaene et al., What is consciousness, and could machines have it, Science 2017

➢ Our goal: architecture that resembles the C1 functionality in terms of information reusability

System 1 and System 2 AI

System 1

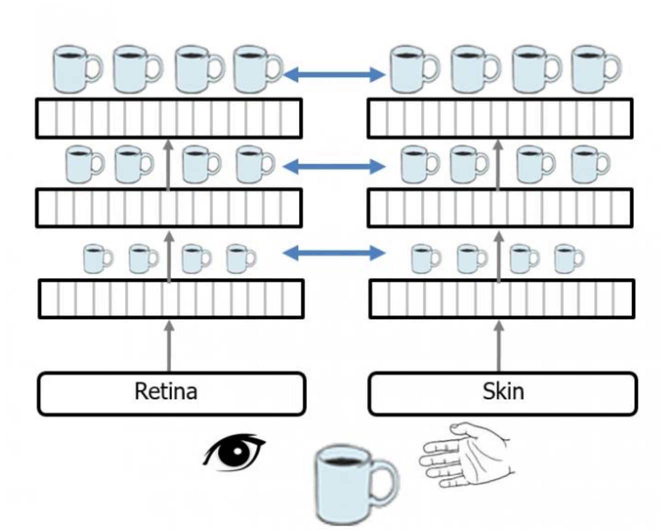
- Intuitive and fast
- Without explanation

→ Monolithic neural networks
CNNs, RNNs, Transformers …

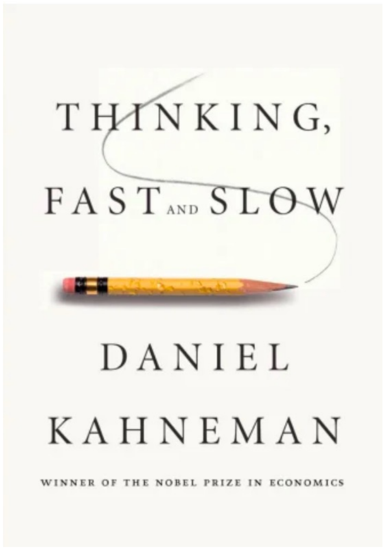
System 2

- Explicit and slow
- Logical reasoning and planning

→ Neural coordination
Localized Learning, Meta Learning…



Thousand Brains Theory, Jeff Hawkins 2018



System 1 and System 2 AI

How to bridge the gap?

System 1

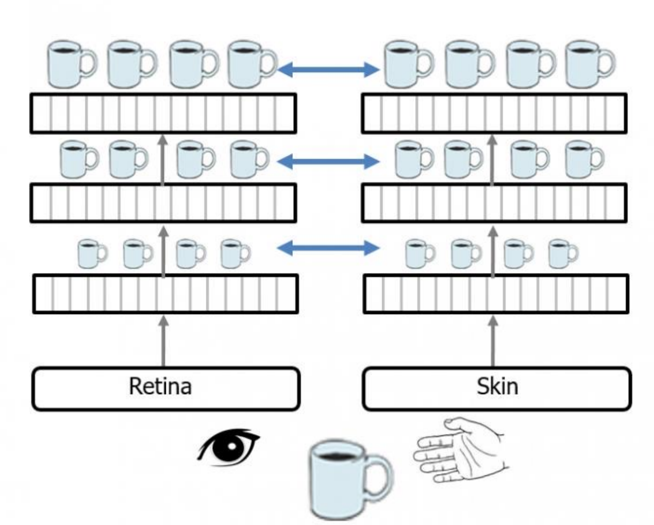
- Intuitive and fast
- Without explanation

→ Monolithic neural networks
CNNs, RNNs, Transformers …

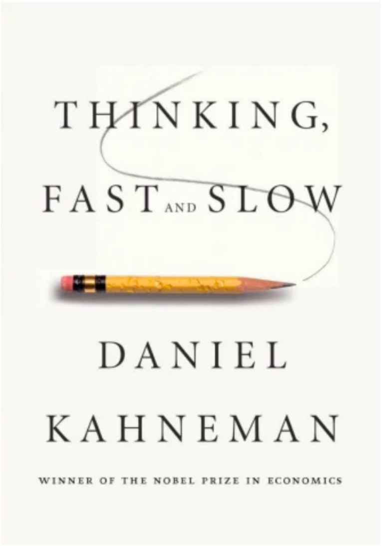
System 2

- Explicit and slow
- Logical reasoning and planning

→ Neural coordination
Localized Learning, Meta Learning…

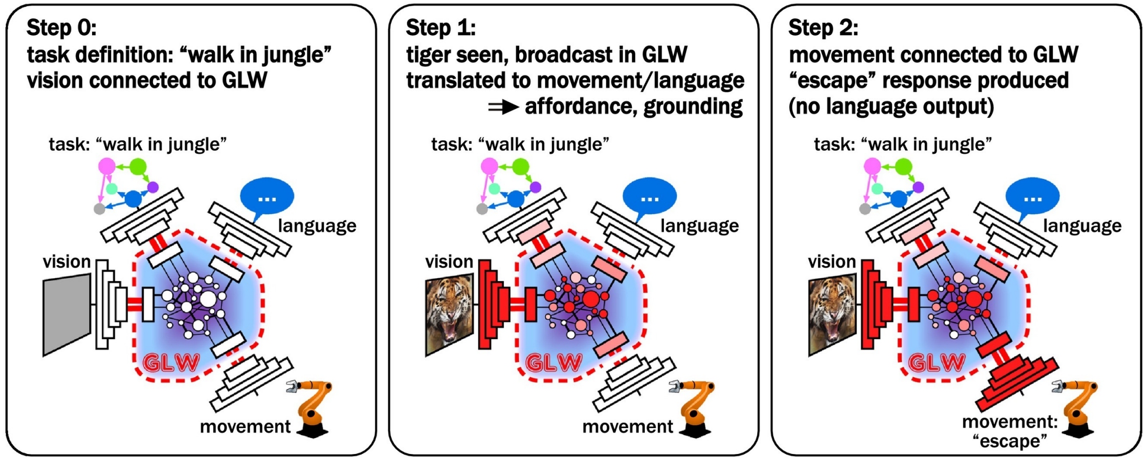


Thousand Brains Theory, Jeff Hawkins 2018

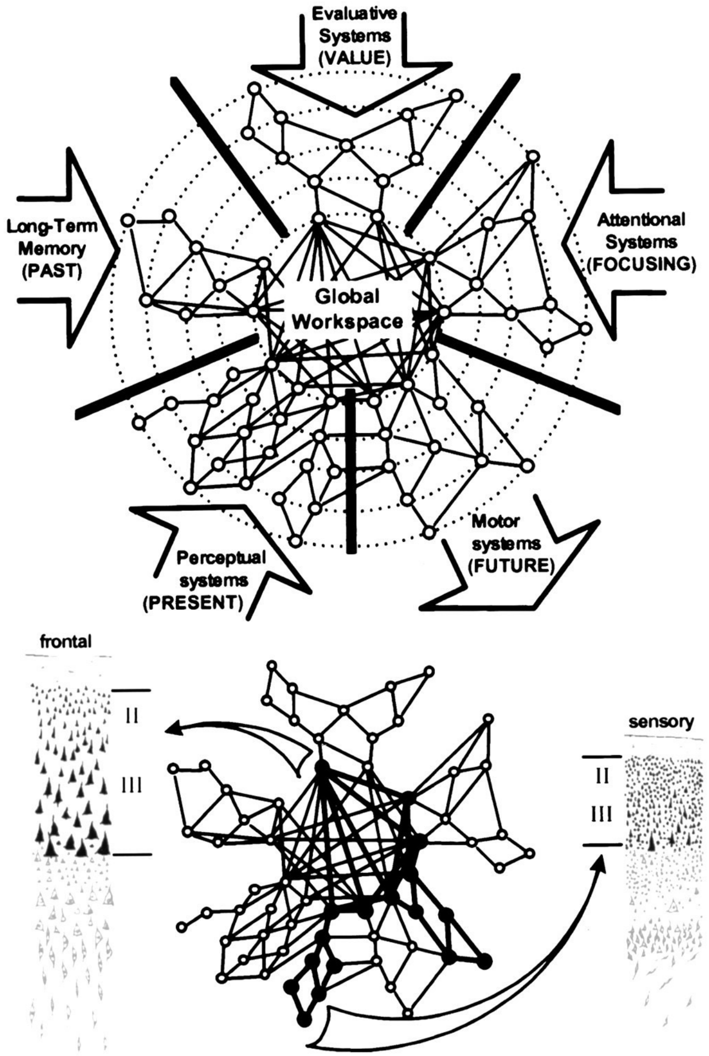


Grounding of Global Workspace Theory for C1 functionality

- A collection of specialized modules
- Guided attention in a limited capacity workspace with a communication bottleneck
- States are conscious when they are broadcast to many modules through the workspace
- Modules compete to share info for better efficiency



[Kanai, 2021]



[Baars, 1988; Dehaene, 1998]

How to implement GW in artificial systems

Global Learning

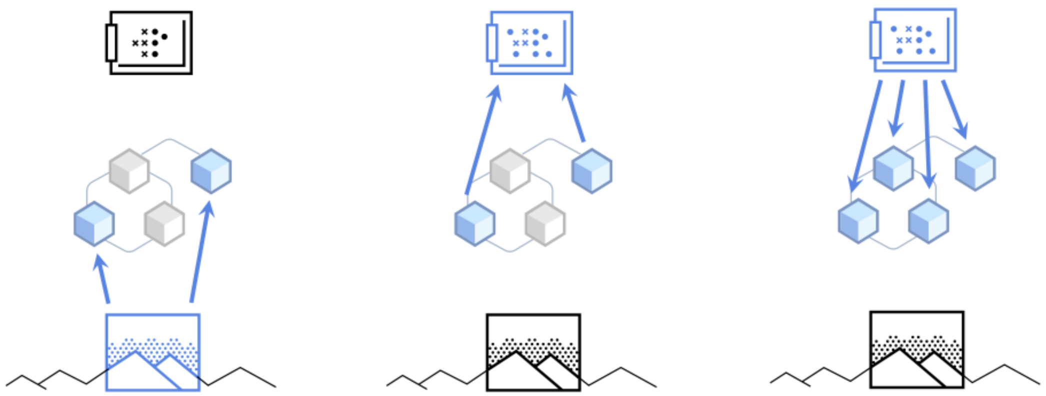
- Learns patterns and relationships over the entire dataset
- Ideal for capturing general trends and insights
- Slow and less suitable for generalization

Localized Learning

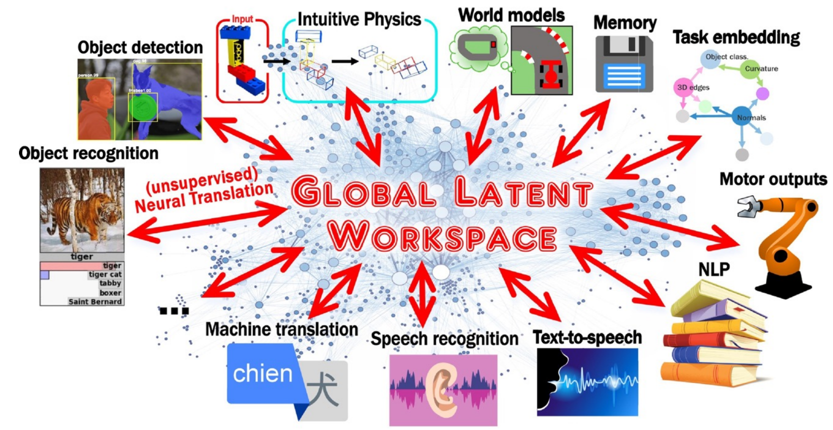
- Learns patterns and features in a restricted context of data for specialized tasks
- Faster convergence on local patterns
- 1) Modules are trained on independent tasks or 2) jointly trained end-to-end, from which module specialization naturally emerges

➢ A collection of specialized modules which can perform tasks in parallel

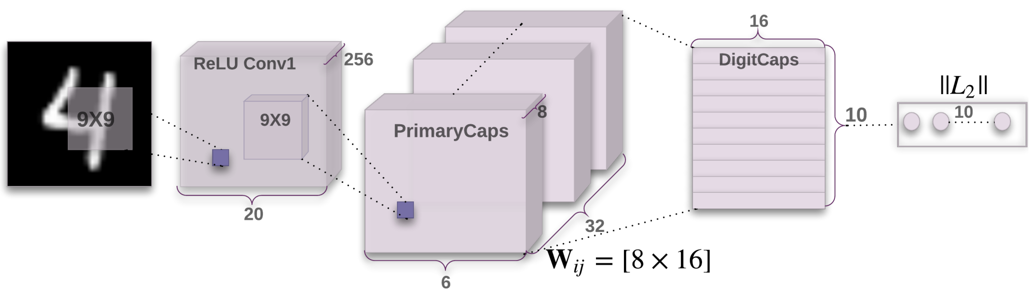
Localized Learning



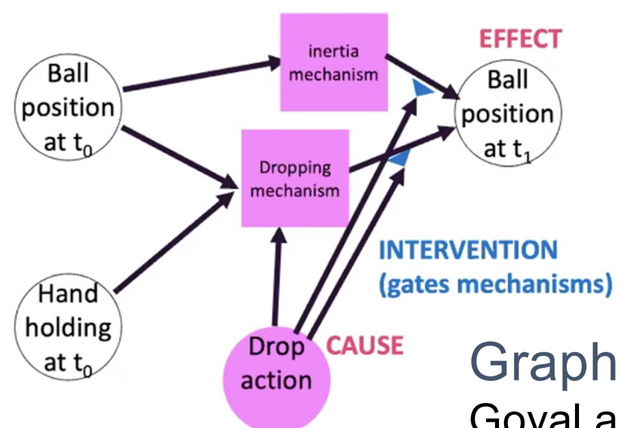
Coordination in Shared Space
Bengio et al., ICLR’22



Global Latent Workspace
Kanai et al., Trends in Neurosciences 2021



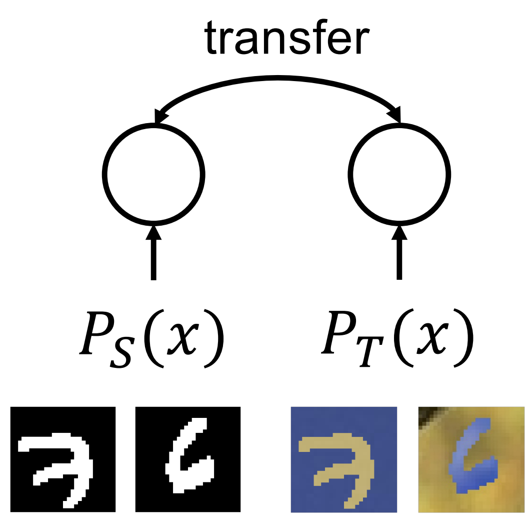
Capsule Networks
Hinton et al., NeurIPS’17



Graph-structured causality
Goyal and Bengio, arXiv’22

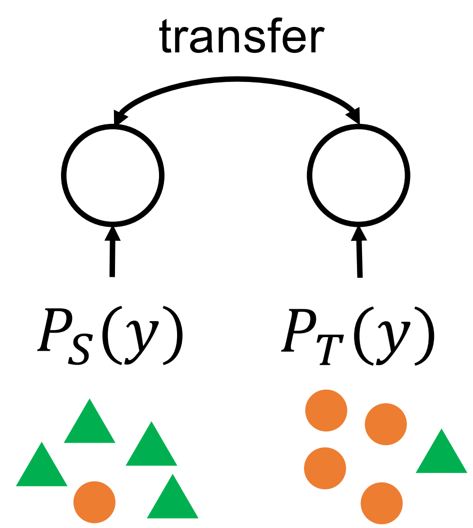
Tractable toy problems of generalization

- Could implementing global workspace improve information alignment among modules, mitigating information loss in knowledge transfer?



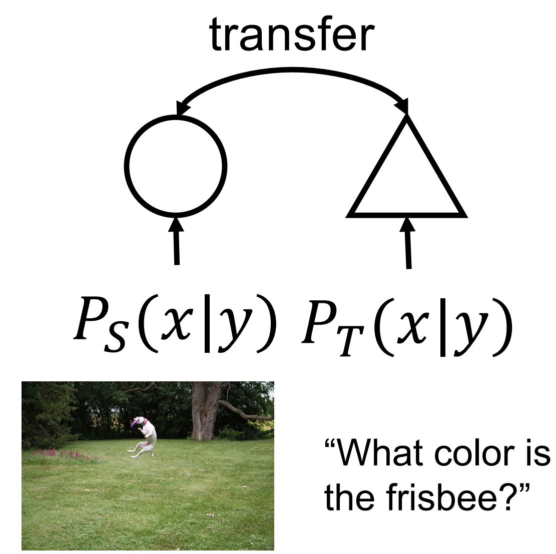
Covariate shift

$P_S(y|x) = P_T(y|x)$,
but $P_S(x) \neq P_T(x)$



Prior probability shift

$P_S(y|x) = P_T(y|x)$, but $P_S(y) \neq P_T(y)$



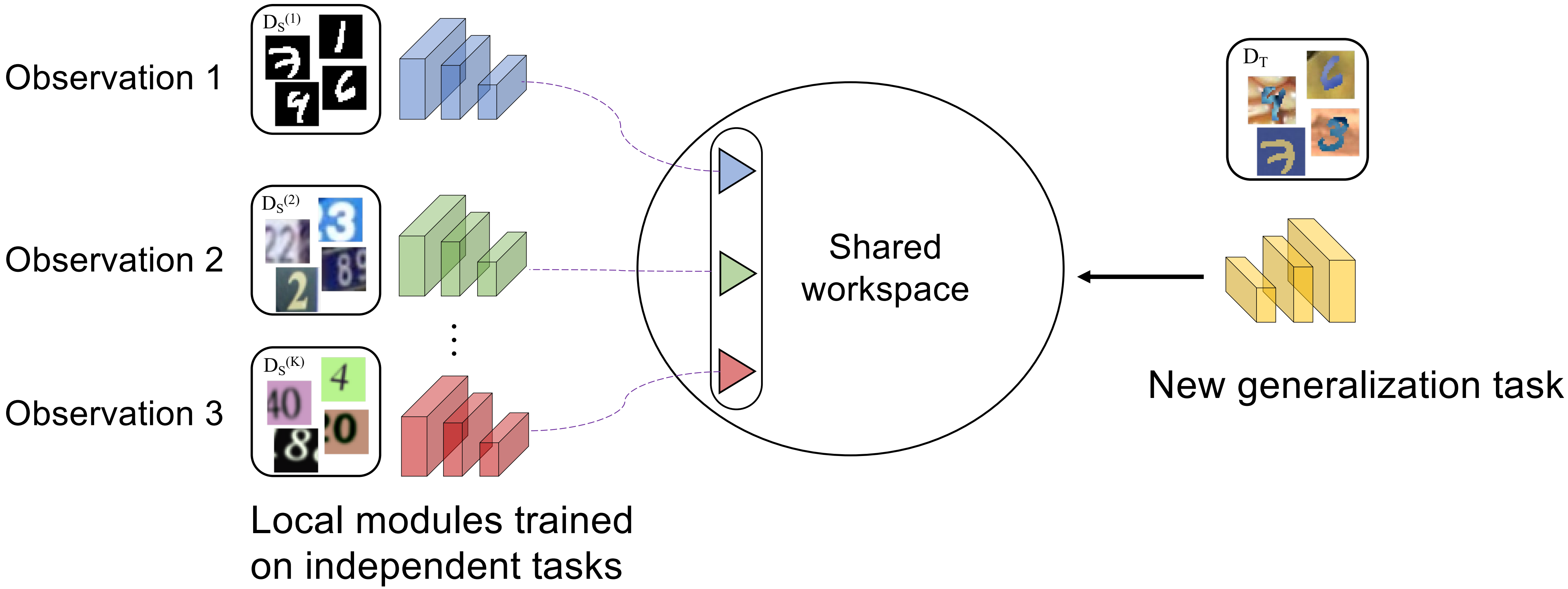
Concept shift

$P_S(x|y) \neq P_T(x|y)$,
but $P_S(y) = P_T(y)$

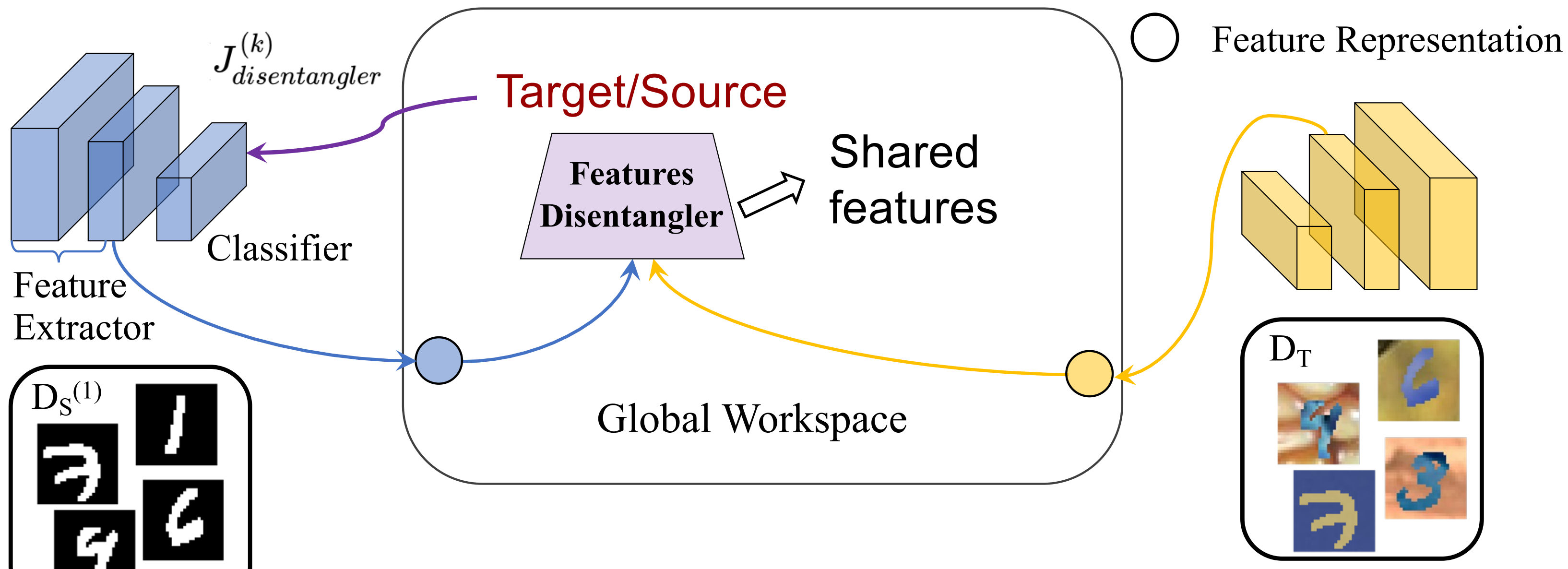
Meta Learning in Decentralized Neural Networks: Towards More General AI, AAAI 23

GW Case 1: modules are trained on independent tasks

- Generalization through shared workspace by reusing localized knowledge from various modules



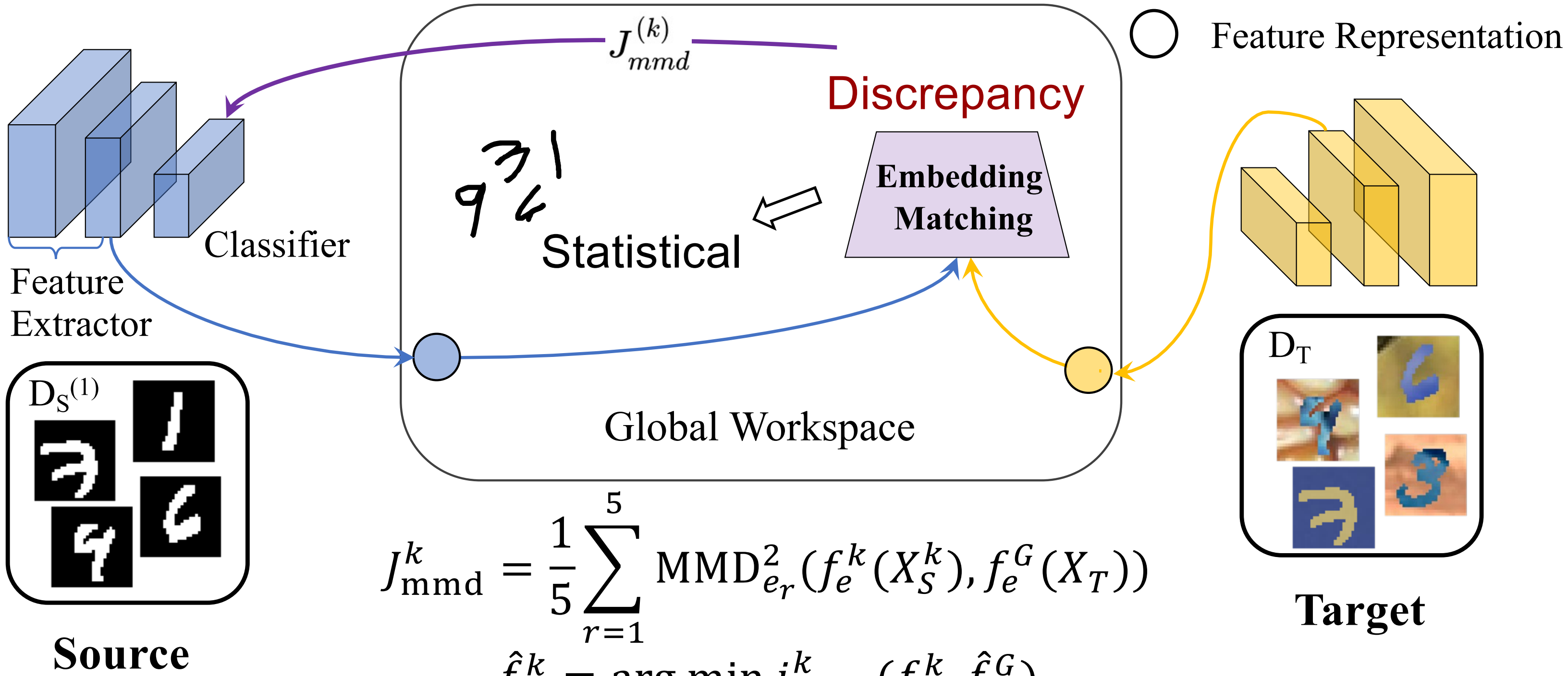
Global features disentangler



$$\hat{f}_e^k = \arg\max_{f_e^k} J^{(k)}_{\text{disentangler}}(\hat{f}_d, f_e^k, \hat{f}_e^G)$$

- Local feature extractors are trained to learn *shared* features
- f_d distinguishes between the source f_e^k and the target f_e^G

Embedding matching



$$J_{\text{mmd}}^{k} = \frac{1}{5}\sum_{r=1}^{5} \text{MMD}_{e_r}^2(f_e^k(X_S^k), f_e^G(X_T))$$

$$\hat{f}_e^k = \arg\min_{f_e^k} j_{\text{mmd}}^k(f_e^k, \hat{f}_e^G)$$

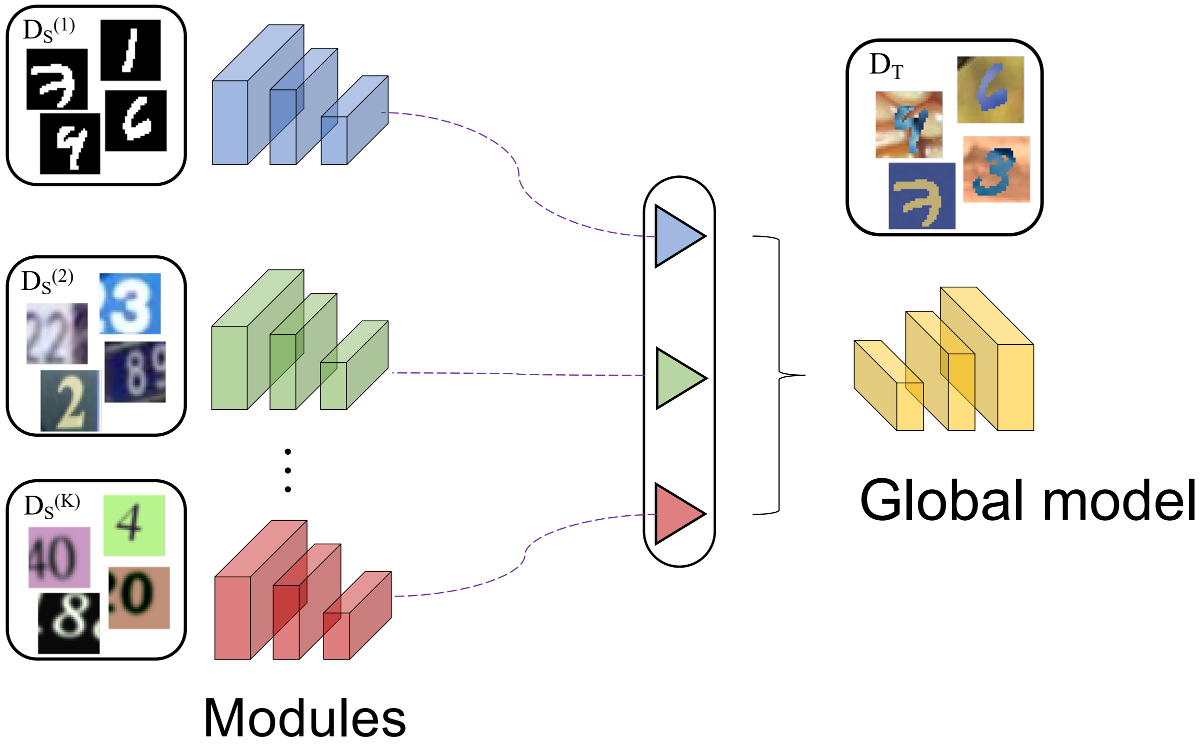
- Common features could be either background noise or objects
- Embedding matching aligns features across multiple observations, which enables the extraction of common objects among these observations, based on a discrepancy loss between f_e^k and f_e^G

Module aggregation

- Alignment process is carried out sequentially for each module
- Aggregate modules to obtain the global model through weight sharing

$$G_{t+1} = G_t + \sum_{k \in K} \frac{N^{(k)}}{\sum_{k \in K} N^{(k)}} (L_{t+1}^{(k)} - G_t)$$

- Broadcast the global model to replace each module

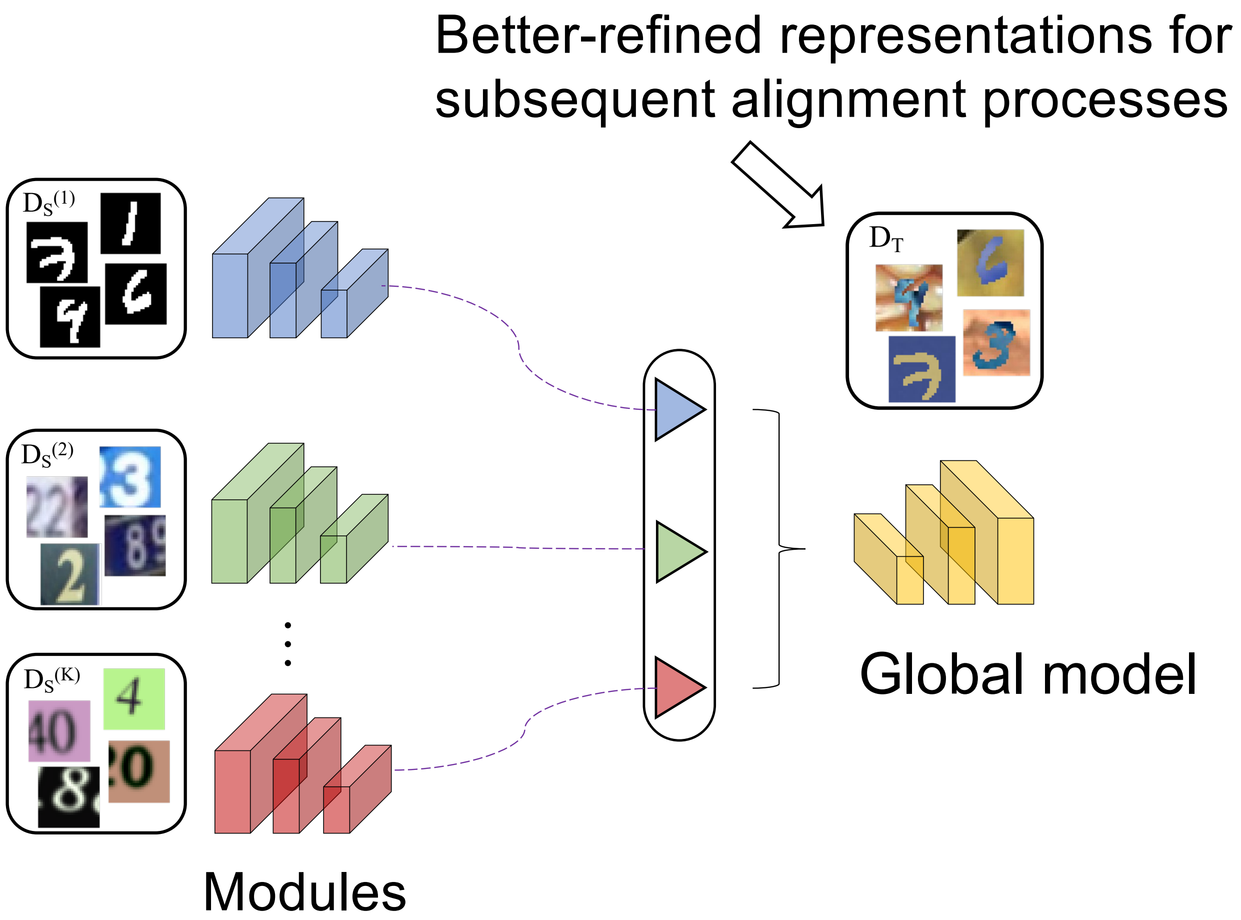


Module aggregation

- Alignment process is carried out sequentially for each module
- Aggregate modules to obtain the global model through weight sharing

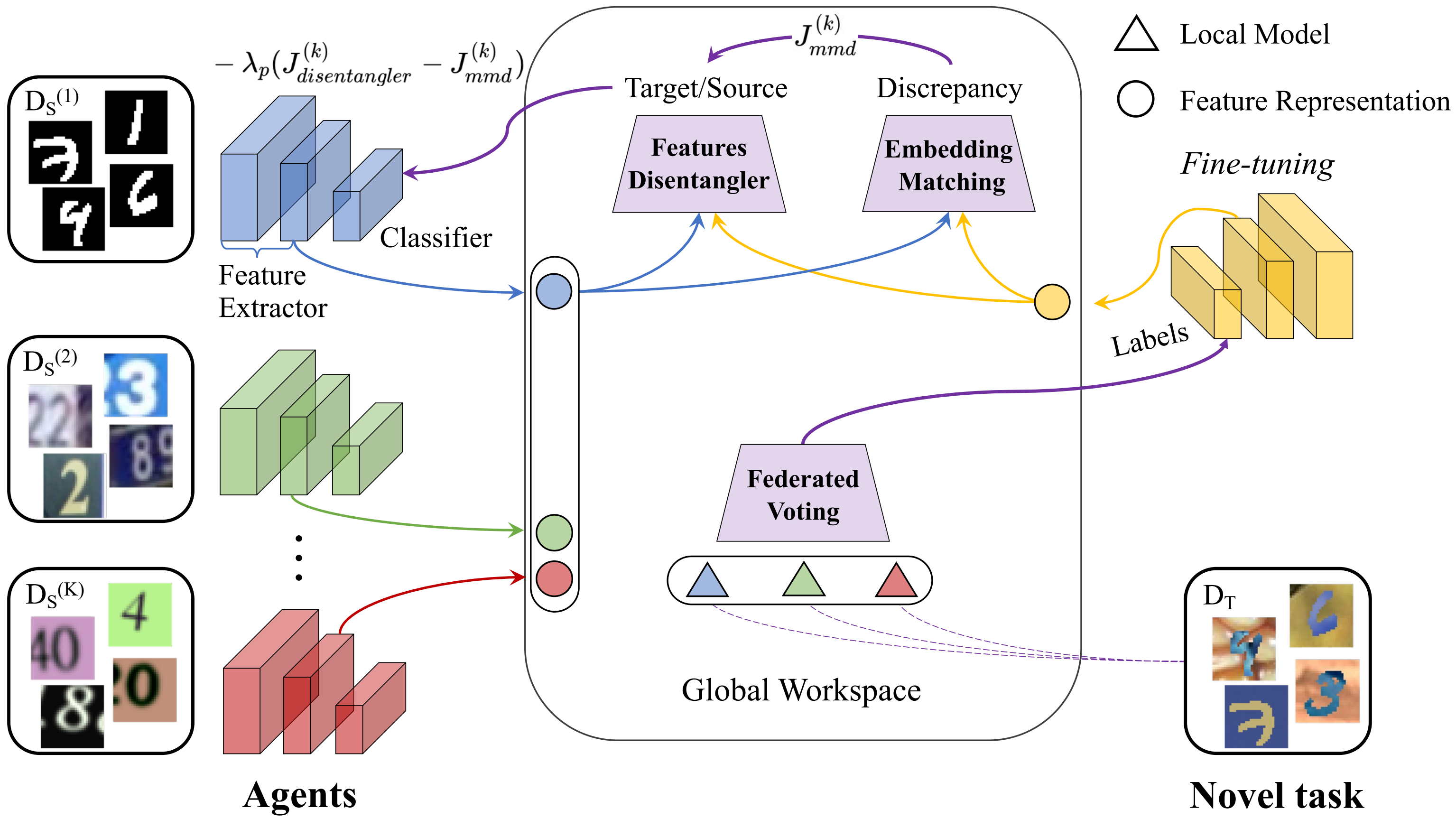
$$G_{t+1} = G_t + \sum_{k \in K} \frac{N^{(k)}}{\sum_{k \in K} N^{(k)}} (L_{t+1}^{(k)} - G_t)$$

- Broadcast the global model to replace each module



GW Case 1: modules are trained on independent tasks

- Generalization through shared workspace by reusing localized knowledge from various modules



Feature Distribution Matching for Federated Domain Generalization, ACML 22

Transfer learning in vision and language tasks

Digit-Five [Ganin, 2015]

Office-Caltech10 [Gong, 2012]

Amazon review [Blitzer, 2007]

DVD: This is a great DVD for all collections (Positive)

Book: This book turns the entire concept of intelligence inside out (Negative)

Electronics: This is perfect for my iPod and keeps it totally secure (Positive)

Kitchen: Simple, straight forward to use, very easy to clean, and durable (Positive)

➢ Classification tasks with an unsupervised approach based on the shared workspace

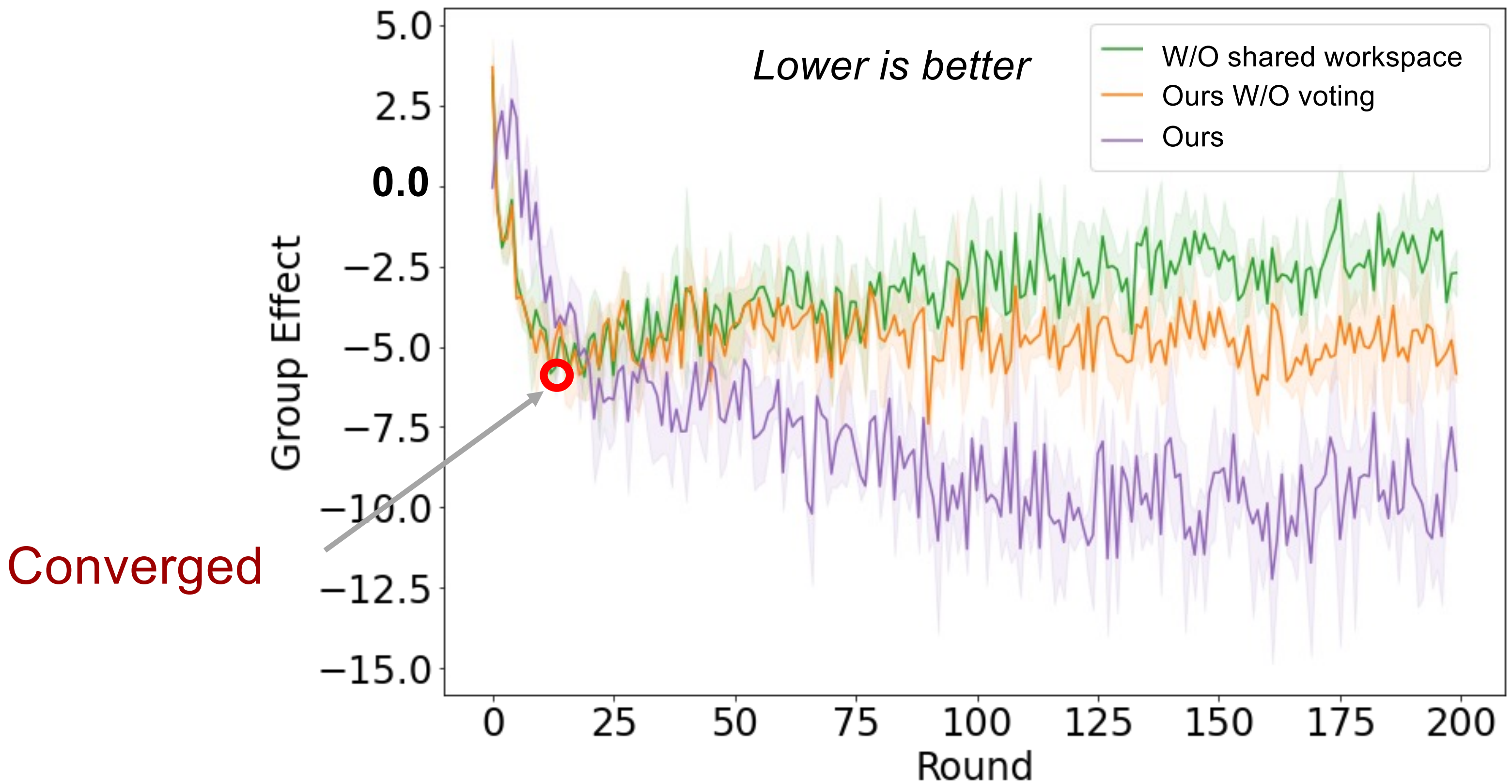
Measuring information loss in the shared workspace

- Target Task Accuracy (TTA)

$$\mathrm{TTA}_f(G_t) = \frac{\sum_{(x,y)\in D_T} \mathbb{1}\{\arg\max_j f(x;G_t)_j = y\}}{|D_T|}$$

- Group Effect (GE)

$$\mathrm{GE}_t = \frac{1}{K}\sum_{k\in\{1,2,\ldots,K\}} \mathrm{TTA}_f(G_t + \Delta_t^{(k)}) - \mathrm{TTA}_f(G_{t+1})$$



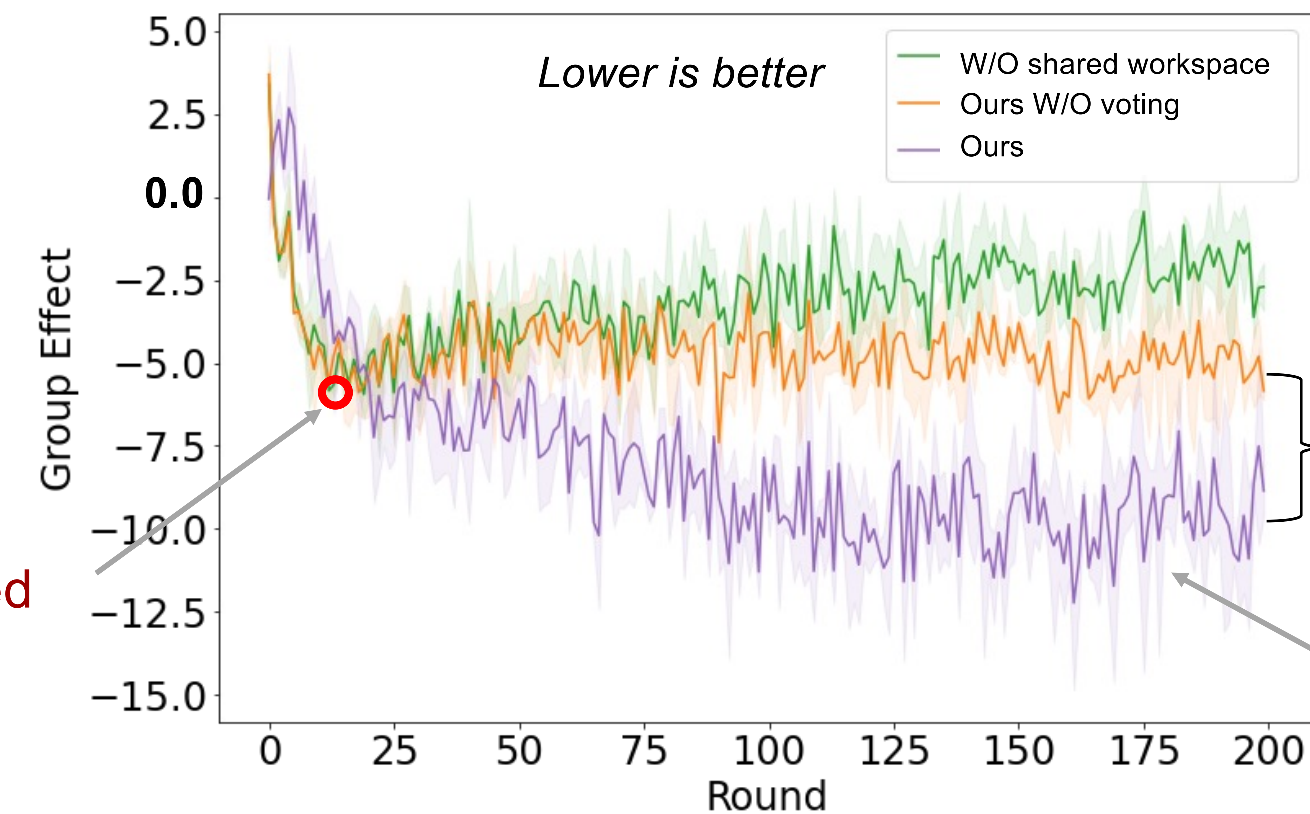
Measuring information loss in the shared workspace

- Target Task Accuracy (TTA)

$$\mathrm{TTA}_f(G_t) = \frac{\sum_{(x,y)\in D_T} \mathbb{1}\{\arg\max_j f(x;G_t)_j = y\}}{|D_T|}$$

- Group Effect (GE)

$$\mathrm{GE}_t = \frac{1}{K}\sum_{k\in\{1,2,\dots,K\}} \mathrm{TTA}_f(G_t + \Delta_t^{(k)}) - \mathrm{TTA}_f(G_{t+1})$$



Converged

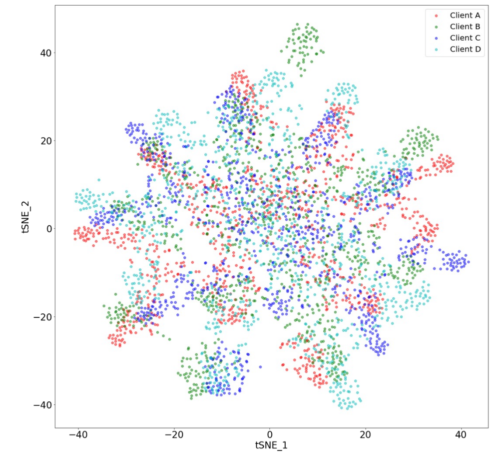
Reduced information loss in the workspace (improved module info transferability)

Late convergence

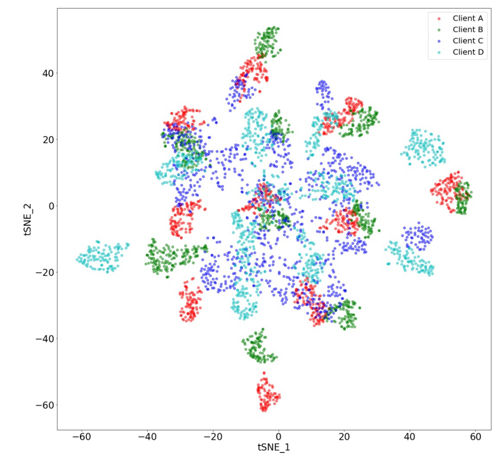
Improved performance and better aligned representations

Models/Tasks Digit Five	→mt	→mm	→up	→sv	→sy	Avg
FedAvg	93.5±0.15	62.5±0.72	90.2±0.37	12.6±0.31	40.9±0.50	59.9
f-DANN	89.7±0.23	70.4±0.69	88.0±0.23	11.9±0.50	43.8±1.04	60.8
f-DAN	93.5±0.26	62.1±0.45	90.2±0.13	12.1±0.56	41.5±0.76	59.9
Voting-S	**93.7**±0.18	63.4±0.28	**92.6**±0.25	14.2±0.99	45.3±0.34	61.8
Voting-L	93.5±0.18	64.8±1.01	92.3±0.21	14.3±0.42	45.6±0.57	62.1
Disentangler + Voting-S	91.8±0.20	71.2±0.40	91.0±0.58	14.4±1.09	48.7±1.19	63.4
Disentangler + Voting-L	92.1±0.16	71.8±0.48	90.9±0.36	15.1±0.91	49.1±1.03	63.8
Disentangler + MK-MMD	90.0±0.49	70.4±0.86	87.5±0.25	12.2±0.70	44.3±1.18	60.9
FedKA-S	91.8±0.19	72.5±0.91	90.6±0.14	**15.2**±0.46	48.9±0.48	63.8
FedKA-L	92.0±0.26	**72.6**±1.03	91.1±0.24	14.8±0.41	**49.2**±0.78	**63.9**

Module representation distribution



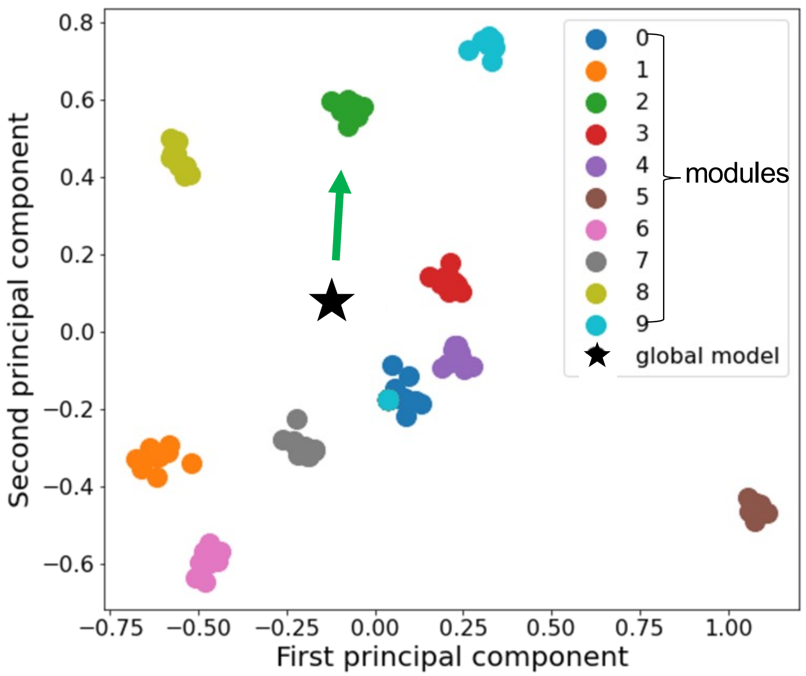
Add shared workspace



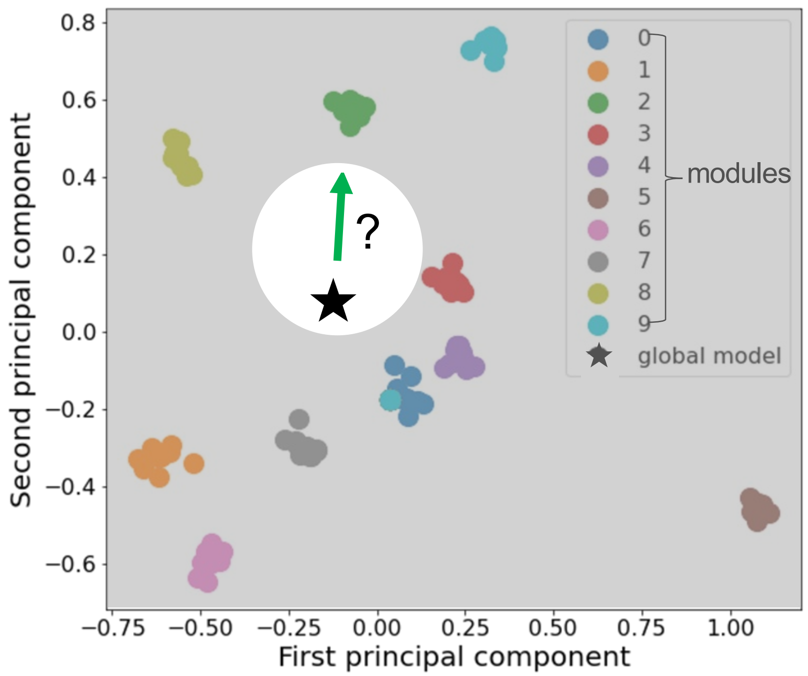
Models/Tasks Office-Caltech	C,D,W→A	A,D,W→C	C,A,W→D	C,D,A→W	Avg
FedAvg	56.4 ±1.23	40.2 ±0.69	28.7±1.21	22.7±1.85	37.0
f-DANN	58.3 ±1.53	40.0 ±1.50	30.7 ±3.59	22.3±1.29	37.8
f-DAN	56.7±0.71	38.7±0.75	30.2±1.64	23.9 ±1.70	37.4
Voting	56.5 ±1.88	40.2 ±0.58	29.8±1.45	**24.1** ±0.69	37.7
Disentangler + Voting	**61.4** ±2.51	**40.4** ±1.01	31.5 ±3.11	23.9 ±1.89	**39.3**
Disentangler + MK-MMD	59.5 ±0.41	37.8±0.93	**32.2** ±3.21	22.3 ±1.00	38.0
FedKA	59.9 ±1.44	39.7±0.81	30.2 ±1.71	23.4 ±1.45	38.3

If tasks of modules are unknown

Ideally: adaptively prioritizing some of the modules to enter the space



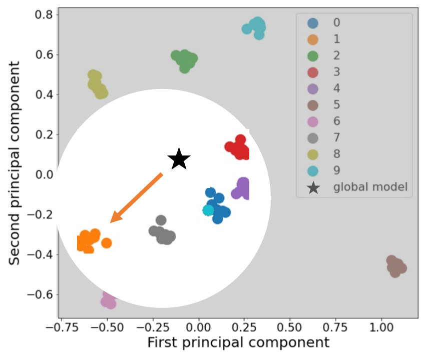
Reality: the task or knowledge distribution of modules is unknown



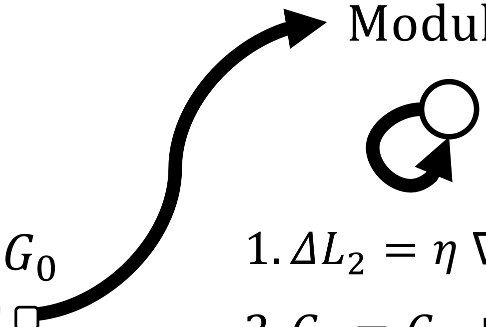
Homogeneous Learning: Self-Attention Decentralized Deep Learning. IEEE Access 2022

Coordination with a Markov decision process

Select a module based on
a learnable policy π



Module_2



G_0

Start

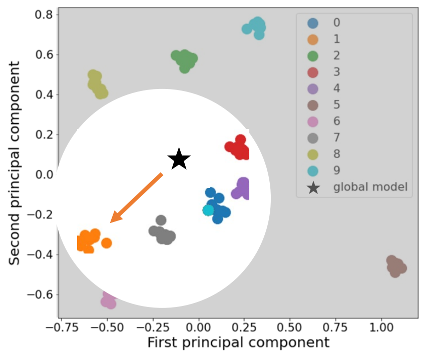
$1.\ \Delta L_2 = \eta\ \nabla J(Y^t, L_2(X^t))$

$2.\ G_1 = G_0 + \Delta L_2$

➢ *Action*: reuse and allow the information sharing from a specific module

Coordination with a Markov decision process

Select a module based on
a learnable policy π



Optimization
objectives

Evaluation

Reward
r_1

Module_2

Module_1

G_0

Start

$1. \, \Delta L_2 = \eta \, \nabla J(Y^t, L_2(X^t))$

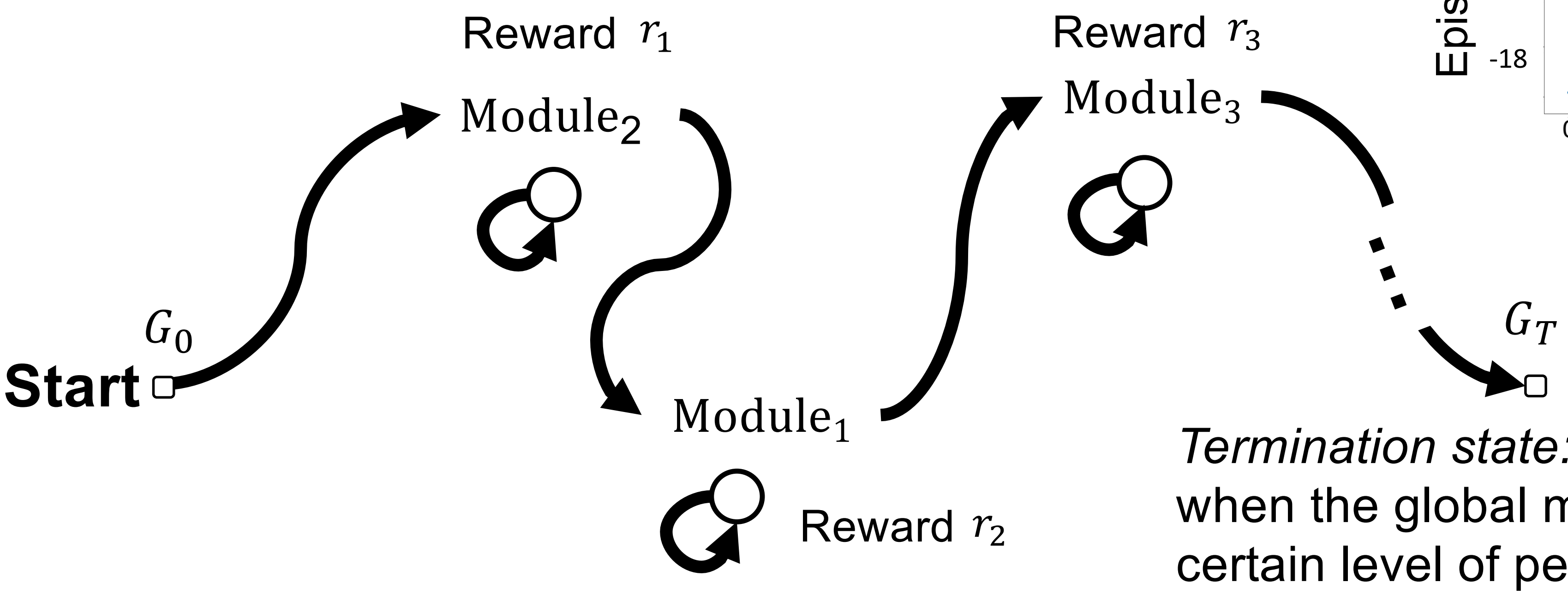
$2. \, G_1 = G_0 + \Delta L_2$

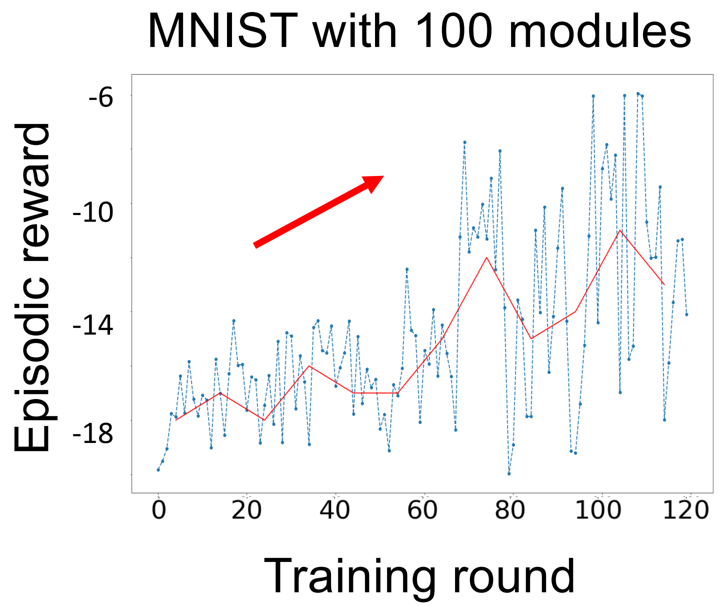
➢ *Action*: reuse and allow the information sharing from a specific module

Coordination with a Markov decision process

- Learn a module selection policy that maximizes the accumulative reward

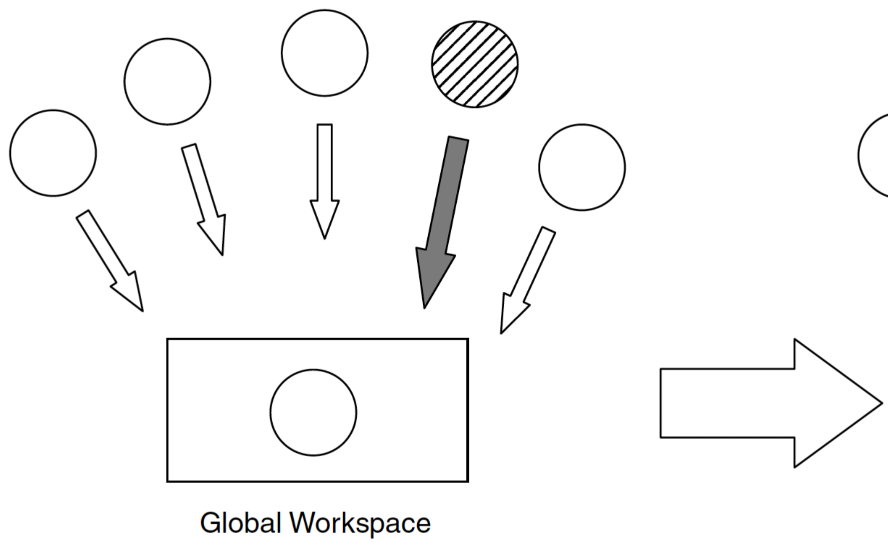
$$\pi_t^* = \arg\max_{\pi_t}(r_t(G_t) + \gamma \cdot \hat{r}_{t+1}(G_{t+1}))$$

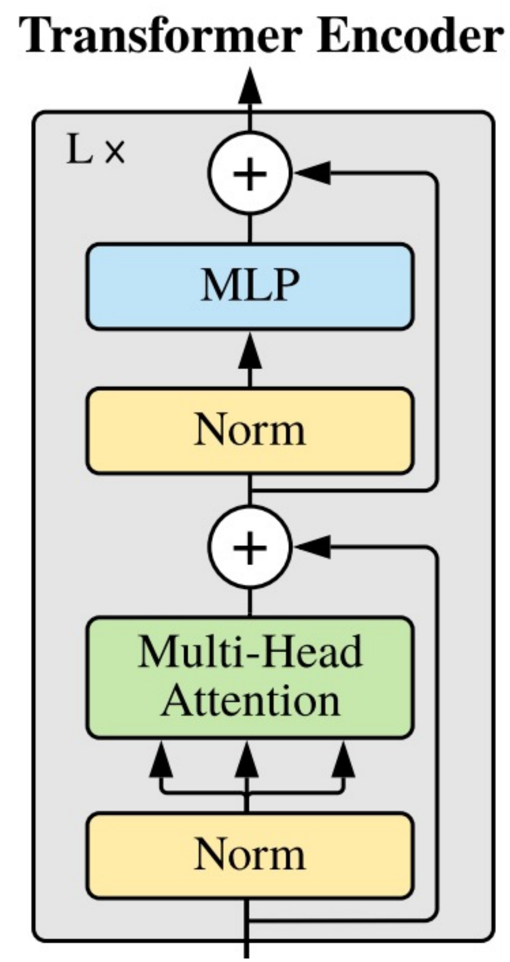




Termination state:
when the global model achieves a certain level of performance or the maximum selection steps are reached

GW Case 2: naturally emerging module specialization

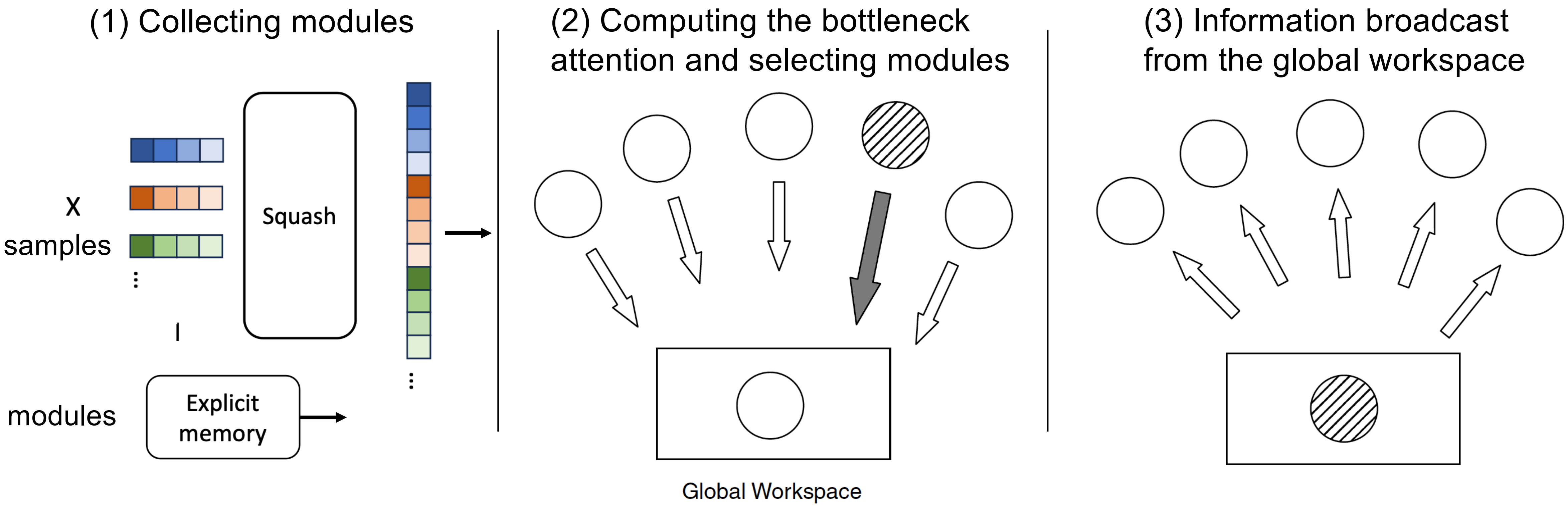




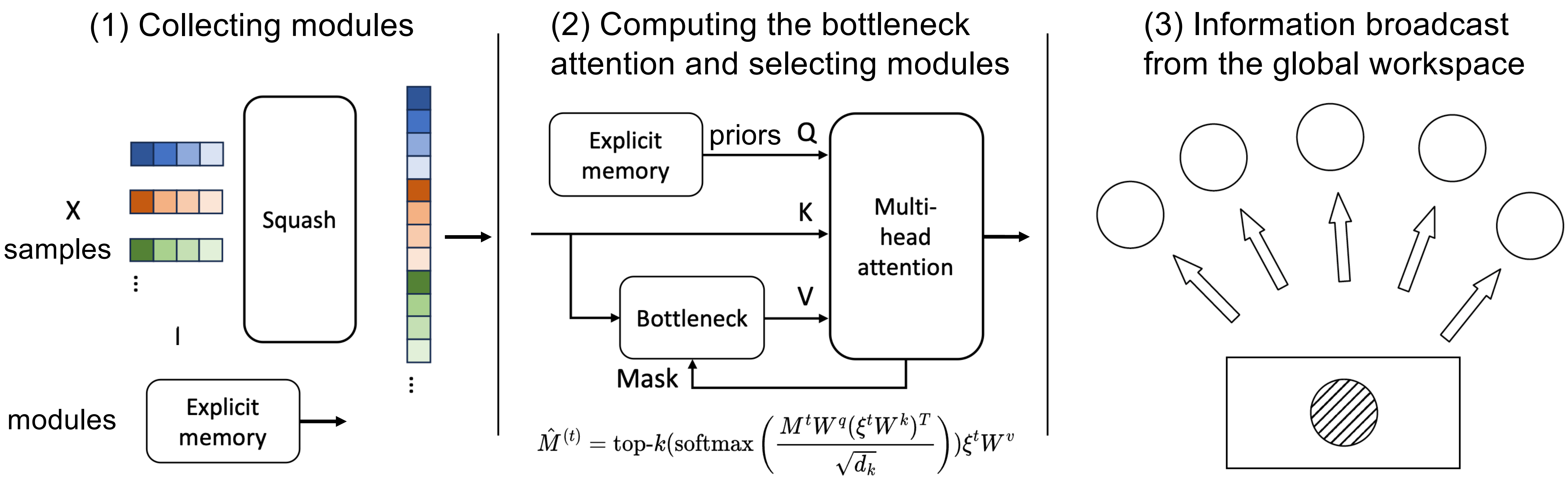
[Dosovitskiy, 2021]

- Pairwise interactions such as the self-attention in Transformers become expensive with scale
- There is an absence of communication bottleneck
- Competition results in naturally emerging module specialization

Inducing global workspace for emerging module specialization

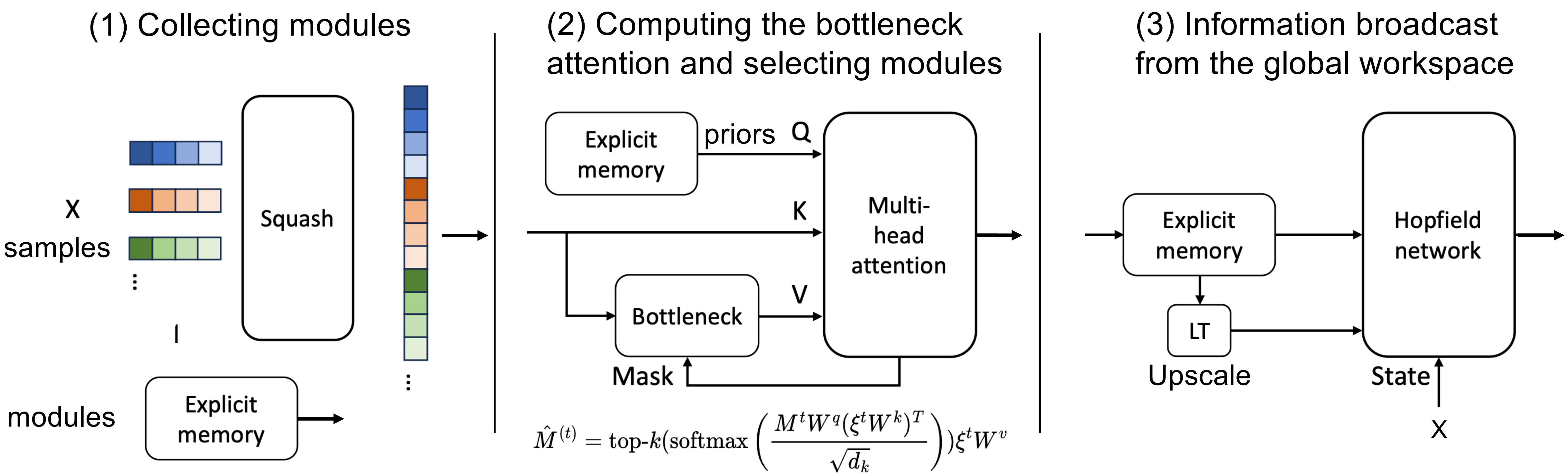


Inducing global workspace for emerging module specialization



➢ Bottleneck allows a few patterns to enter the workspace inducing competition among modules

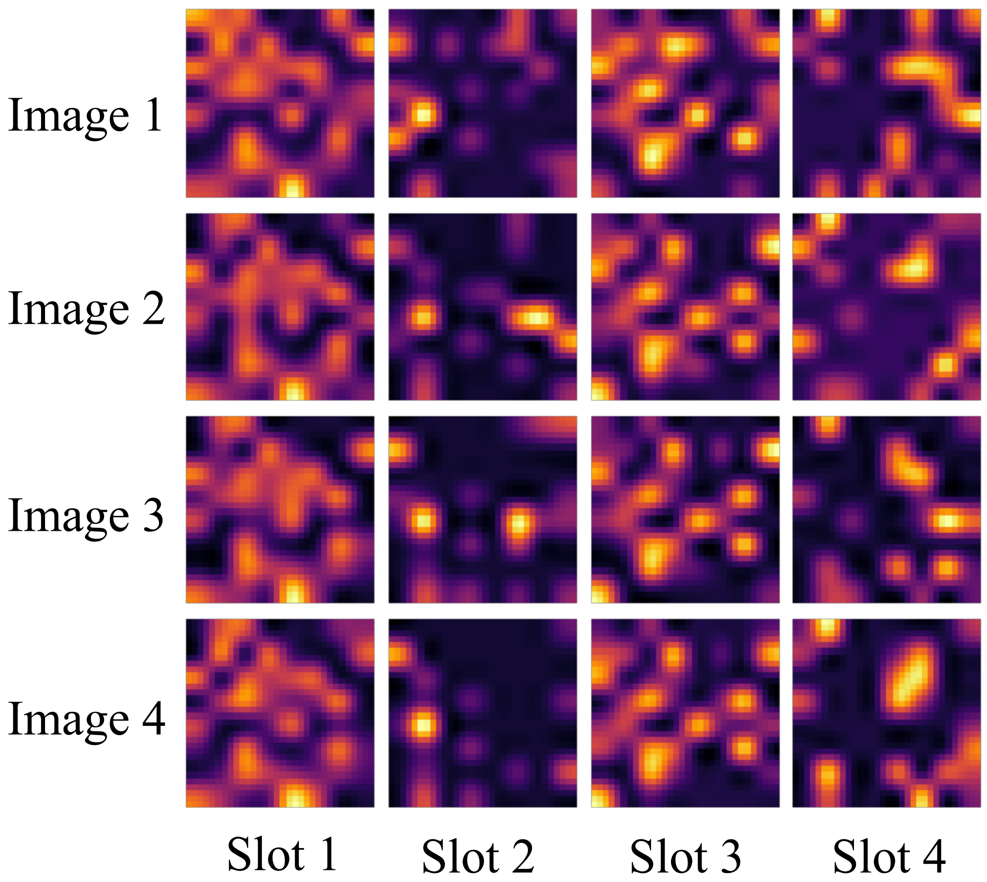
Inducing global workspace for emerging module specialization

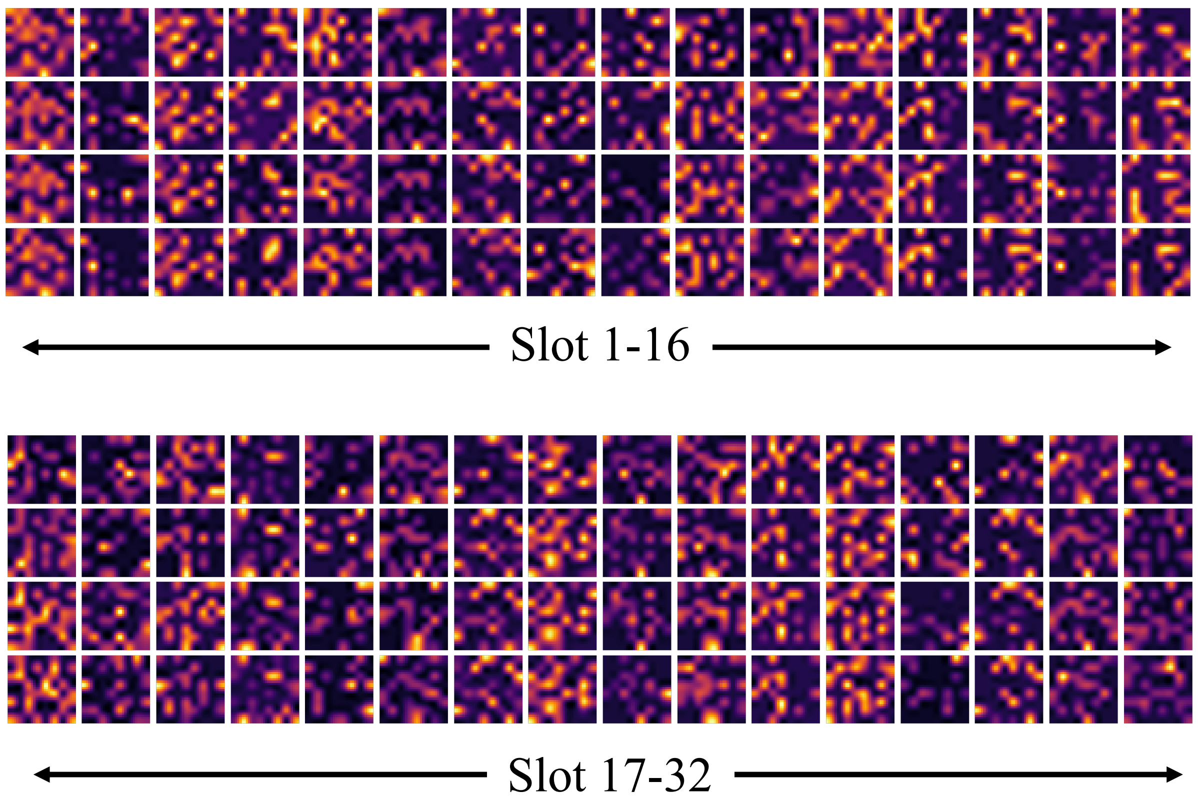


$$\hat{M}^{(t)} = \text{top-}k(\text{softmax}\left(\frac{M^t W^q (\xi^t W^k)^T}{\sqrt{d_k}}\right))\xi^t W^v$$

- Bottleneck allows a few patterns to enter the workspace inducing competition among modules
- Hopfield network uses the learned memory from the bottleneck to reconstruct information that achieves globally lower energy (any neural trajectory that enters an attractor's basin of attraction will converge to that attractor)

Learned bottleneck attention maps in CIFAR10

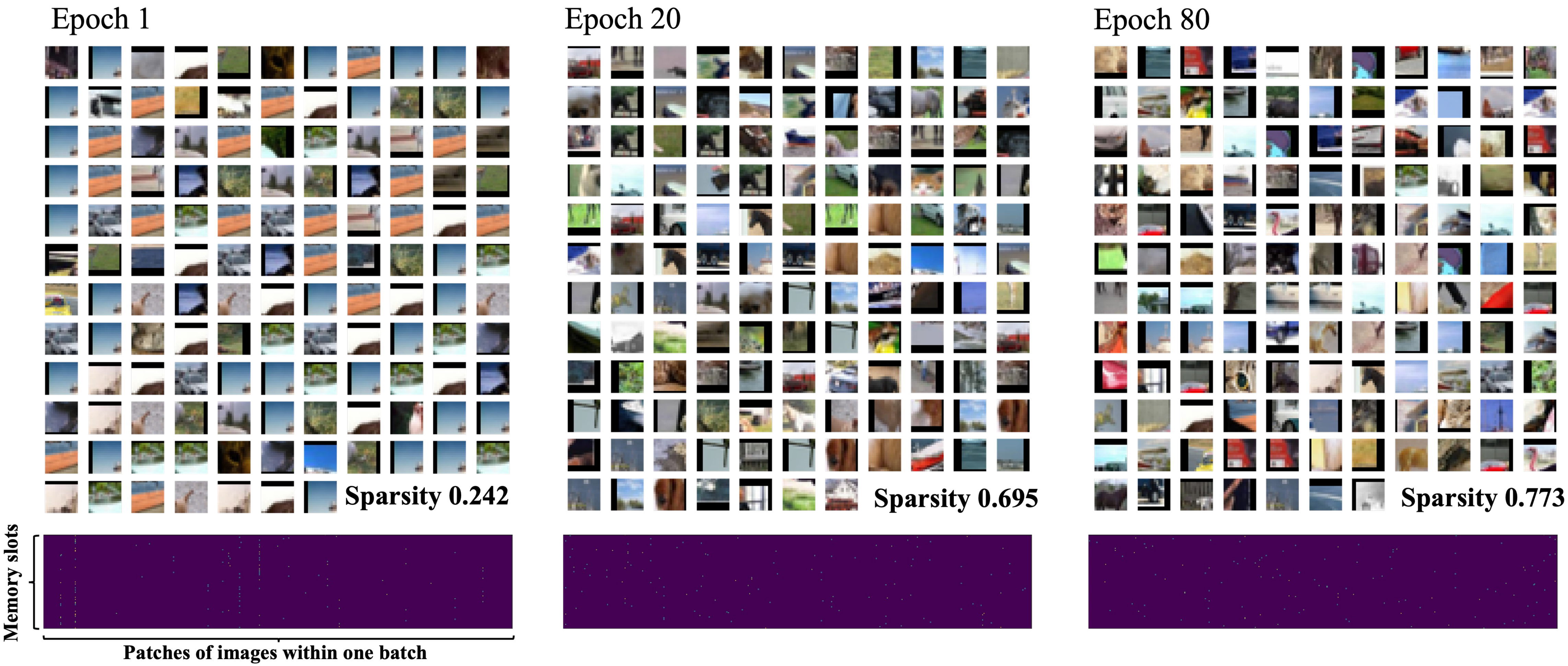




- ➢ Each memory slot learns to attend to a different region of pixels in input images

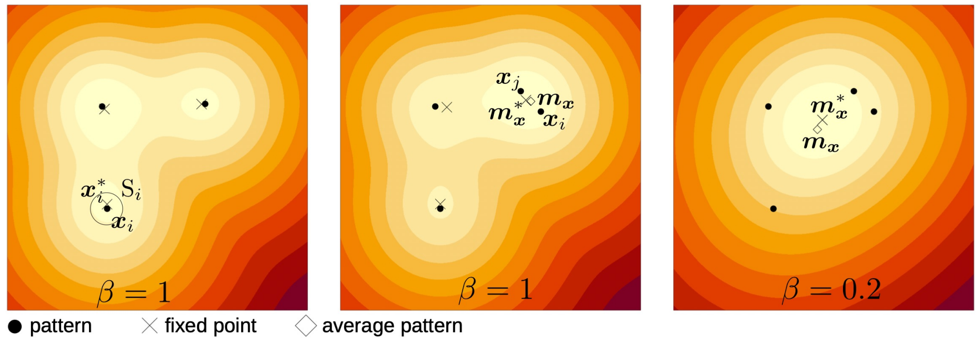
Increased bottleneck attention distribution sparsity

- Selected image patches (modules) by the bottleneck attention



Information retrieval from GW with continuous Hopfield networks

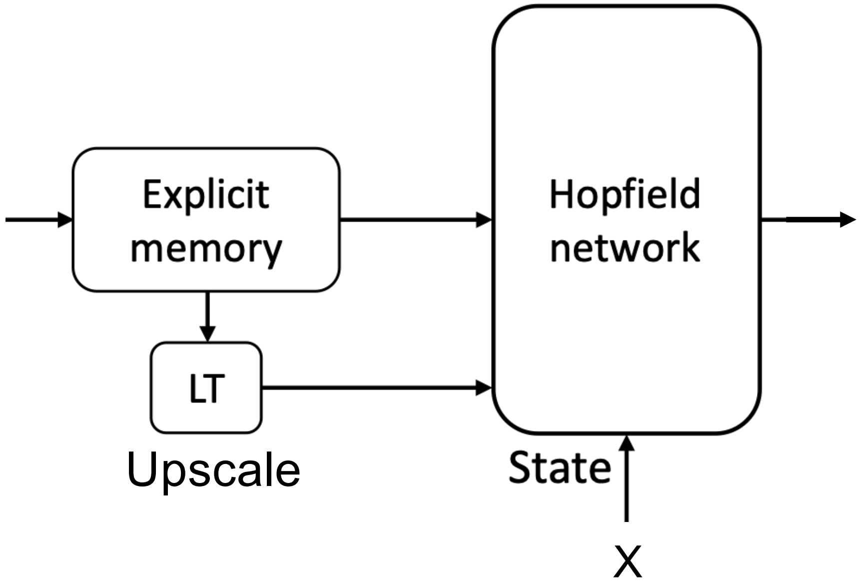
Basin of attraction in continuous Hopfield networks



[Ramsauer, 2021]

fixed points metastable states global fixed point

Methods	CIFAR10	CIFAR100	Triangle	Average
Ours (Base)	**85.44**	**59.10**	99.59	**81.38**
Ours (Small)	83.34	56.30	99.47	79.70
ViT-Base Dosovitskiy et al. (2021)	83.82	57.92	**99.63**	80.46
ViT-Small (Re-impl)	79.53	53.19	99.47	77.40
Perceiver Jaegle et al. (2021)	82.52	52.64	96.78	77.31
Coordination Goyal et al. (2022)	73.42	40.19	97.13	70.25
Bidirectional Mittal et al. (2020)	60.10	31.75	-	45.93
Luna Ma et al. (2021)	47.86	23.38	-	35.62



- Input module states converge to fixed attractor points in the memory of GW
- Emerging module specialization and enhanced performance in small datasets

Conclusions

- Selecting information for global broadcasting and making it flexibly available
- Building an architecture that resembles the C1 functionality by inducing global workspace in conventional ML models
- Modules can be trained on independent tasks or jointly trained end-to-end
- Global workspace helps tackle generalization problems by improving information transferability and encouraging the competition among localized modules
- Inducing global workspace based on the bottleneck attention and Hopfield networks for emerging module specialization

Localized Learning Through the Lens of Global Workspace Theory

Yuwei Sun

University of Tokyo