Meta Learning in Decentralized Neural Networks Towards More General AI

Yuwei Sun

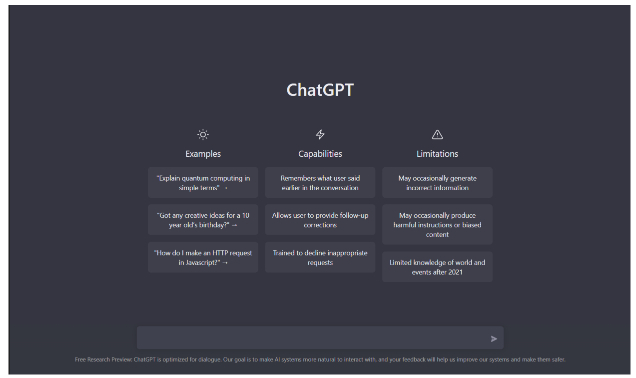
The University of Tokyo

RIKEN

The state of deep learning

Go

Large language model



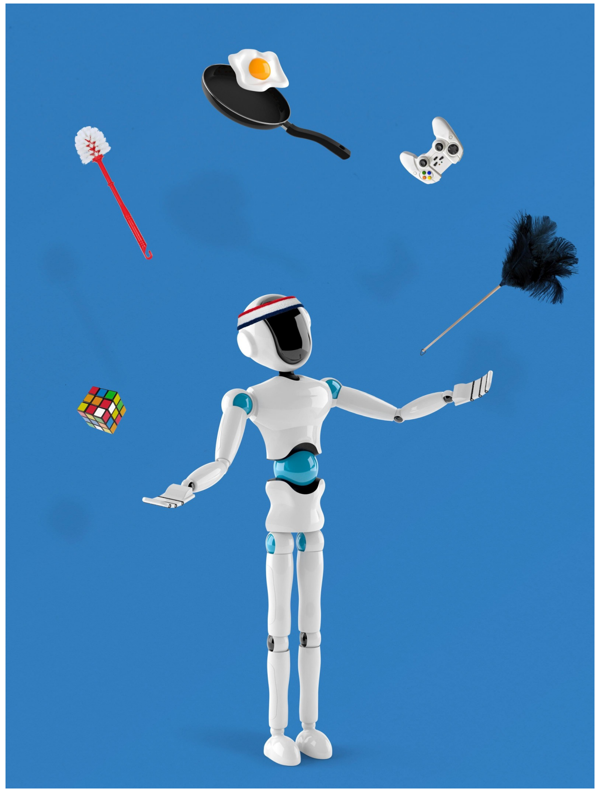
Self-driving car

Text to video

Protein folding

Out-of-distribution generalization

- Training to test
- Distribution difference
- Undesired performance with unseen data

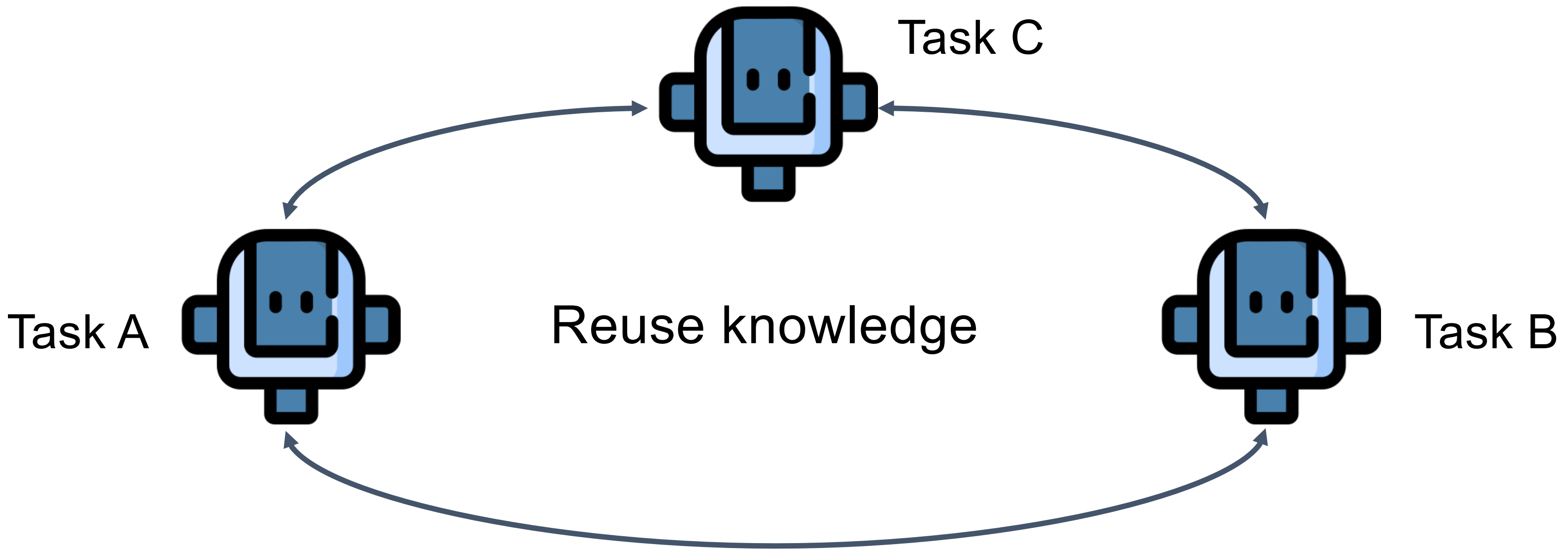


ACM

Beyond the training distribution

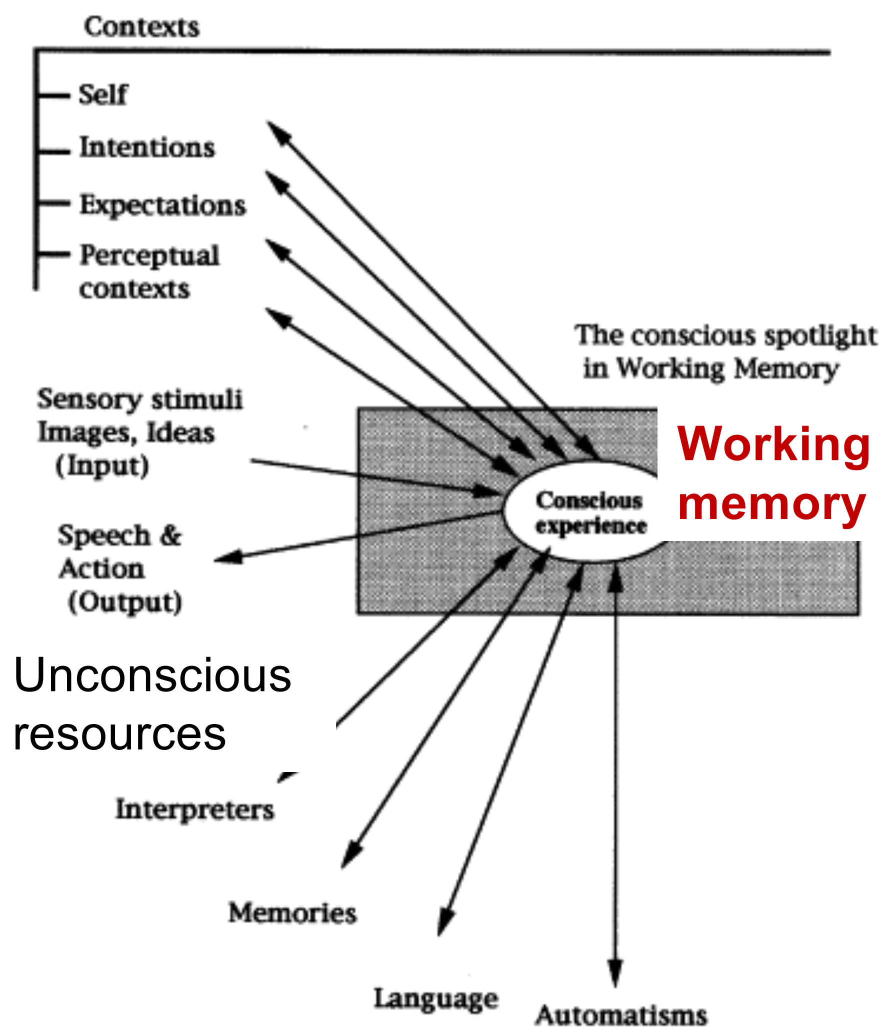
Why meta learning?

- Learning different perspectives with multiple agents (models)
- Meta learning selects and combines learning algorithms to tackle a new task [Pratt, 1991]
- Extract information from past experiences and tasks
- Reusable features and models



Cooperating and competing neural networks

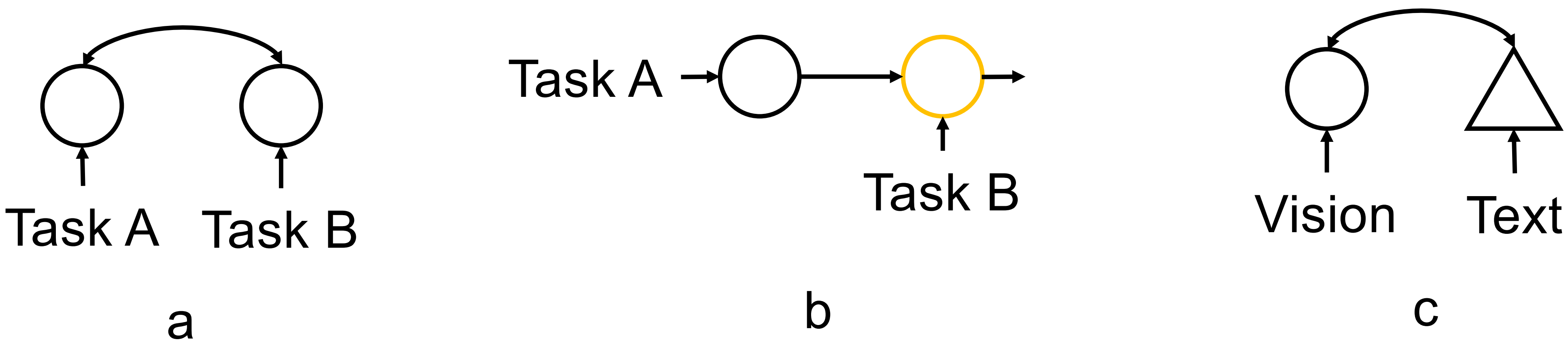
- Global workspace theory [Baars, 1988]
- Cooperating and competing neural network models
- Specialized processors
- Selection and reuse of processors



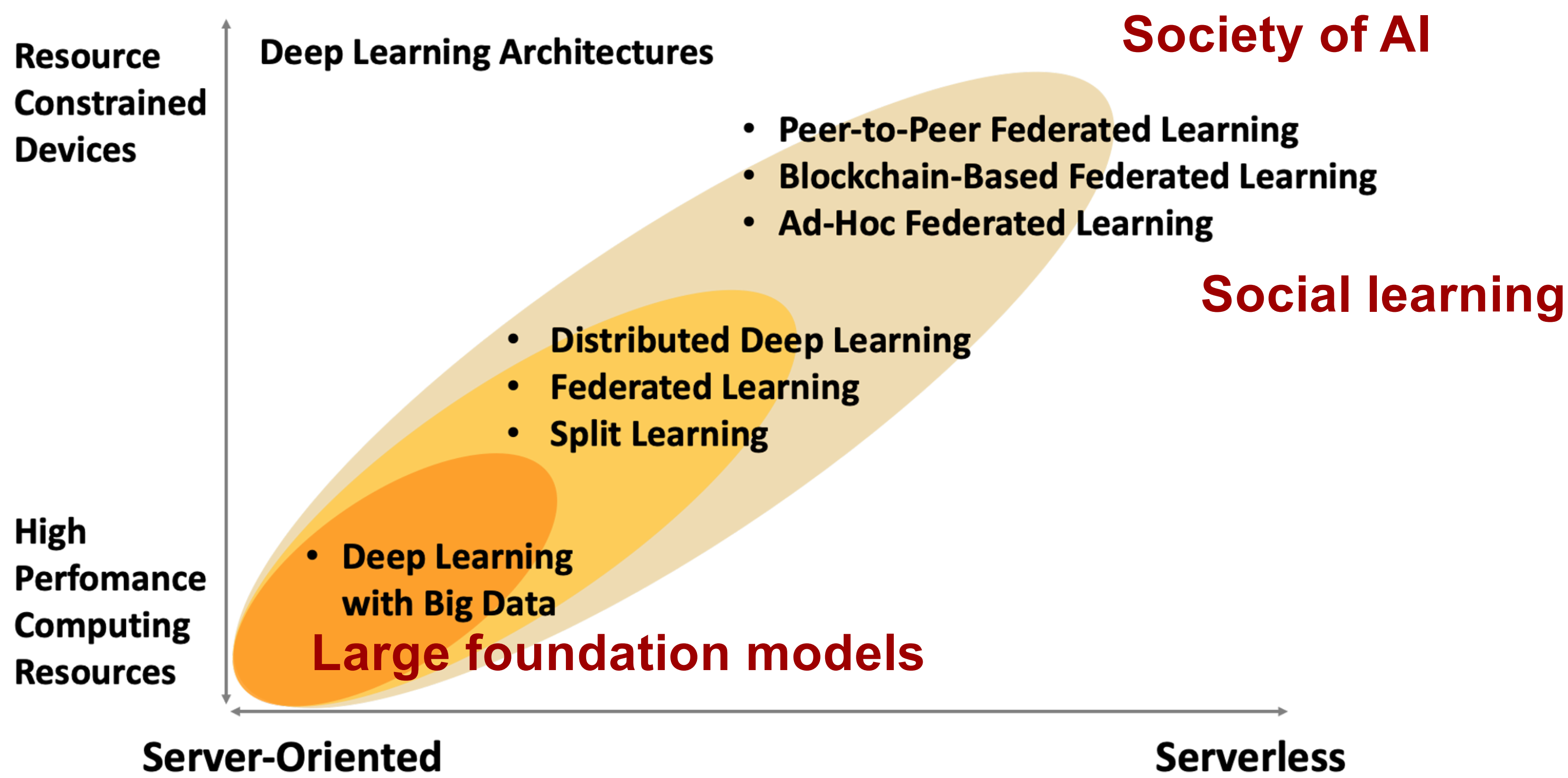
Reusable knowledge representation learning

- Improve neural networks generalization through knowledge transfer

- Discussion on replica neural networks (a), hierarchy of neural networks (b), and multi-modal models (c) as key approaches



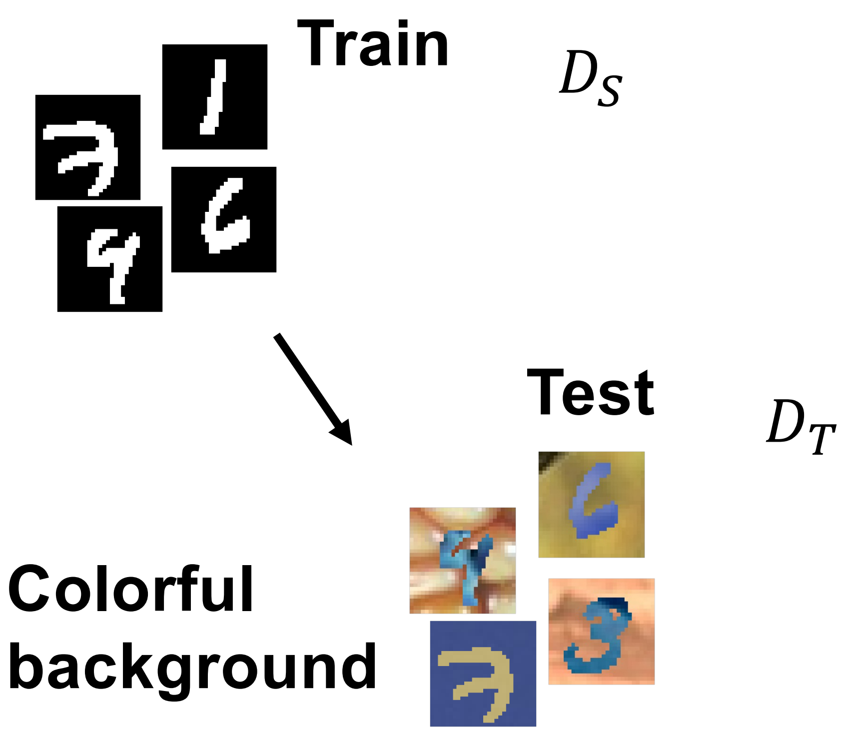
Replica neural networks in multi-agent settings

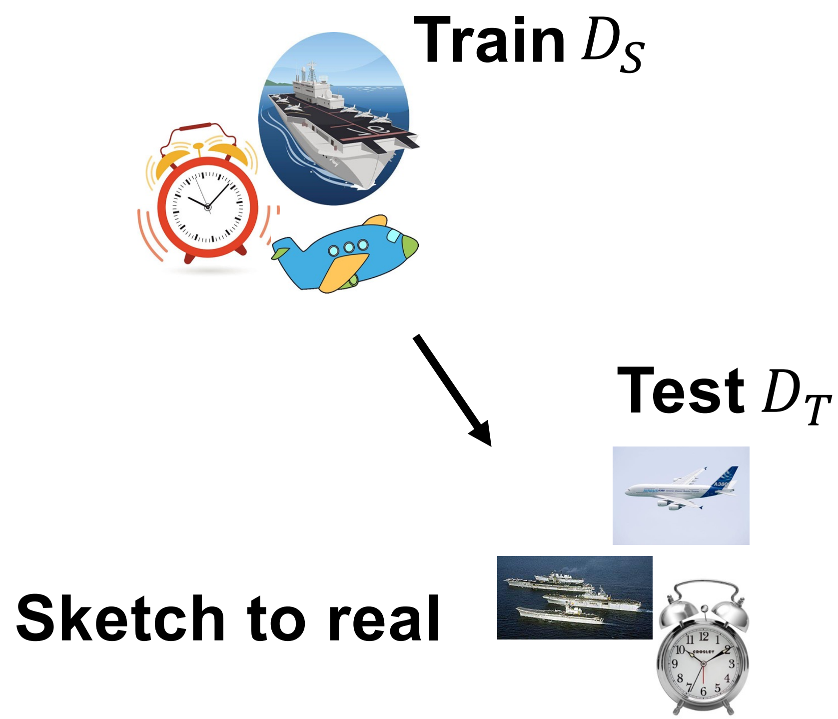


Yuwei Sun, Hideya Ochiai, Hiroshi Esaki. Decentralized Deep Learning for Multi-Access Edge Computing: A Survey. IEEE Transactions on Artificial Intelligence. 2022.

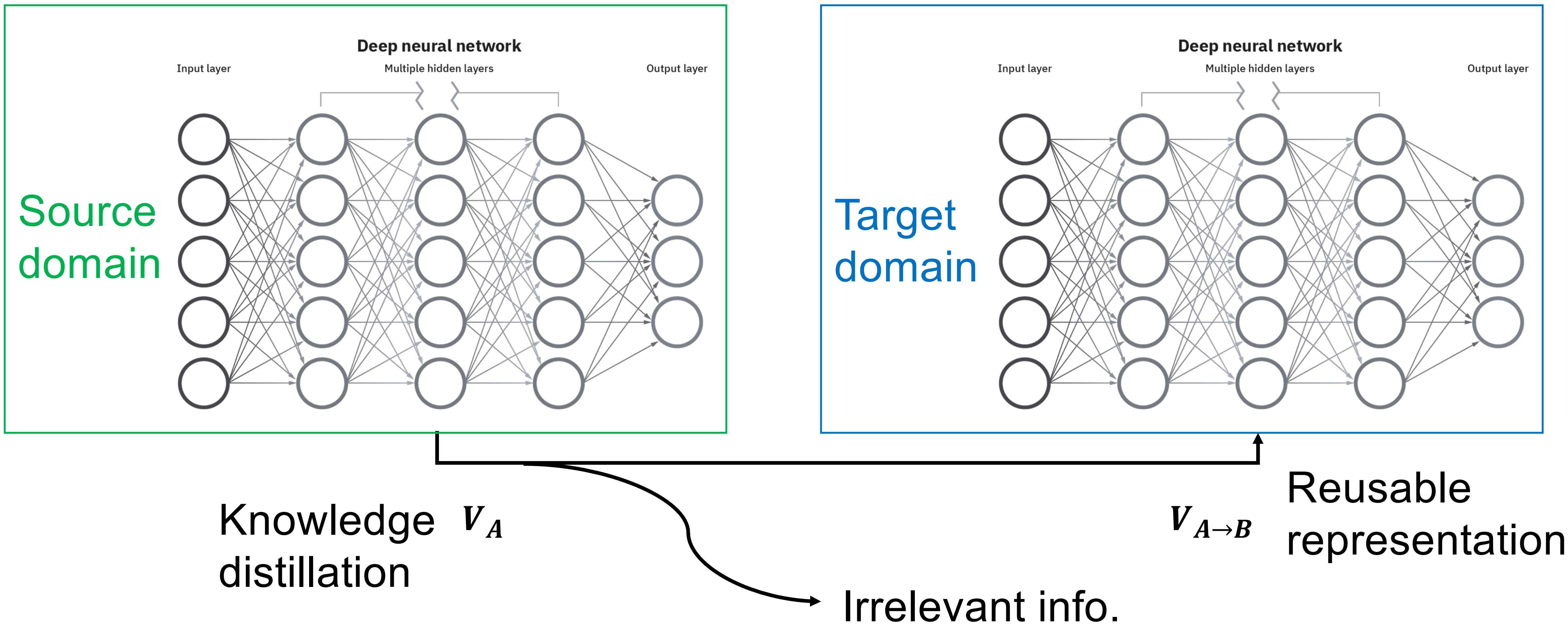
Distributional shift between train and test dataset

- Good in-distribution performance
- Struggle in out-of-distribution (OOD) settings
- Given $D_S = (X_S, Y_S)$ and $D_T = (X_T)$, find $P(Y_T|X_T)$

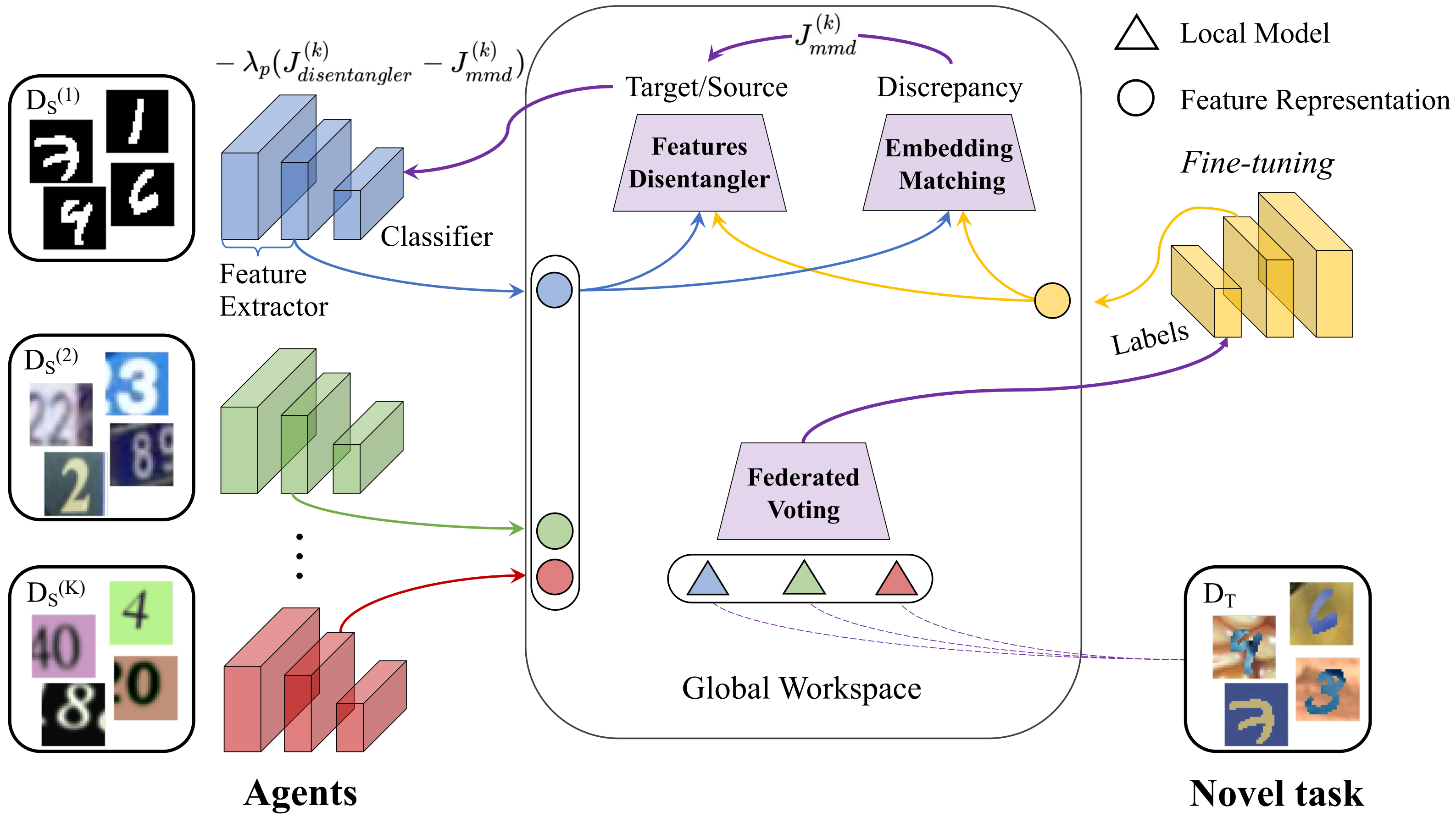




Reusable representation sharing



Feature distribution matching for federated domain generalization



Yuwei Sun, Ng Chong, Hideya Ochiai, Feature Distribution Matching for Federated Domain Generalization, ACML 2022.

Datasets

Digit-Five [Ganin, 2015]

Office-Caltech10 [Gong, 2012]

Amazon review [Blitzer, 2007]

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

DVD: This is a great DVD for all collections

Electronics: This is perfect for my iPod and keeps it totally secure while driving

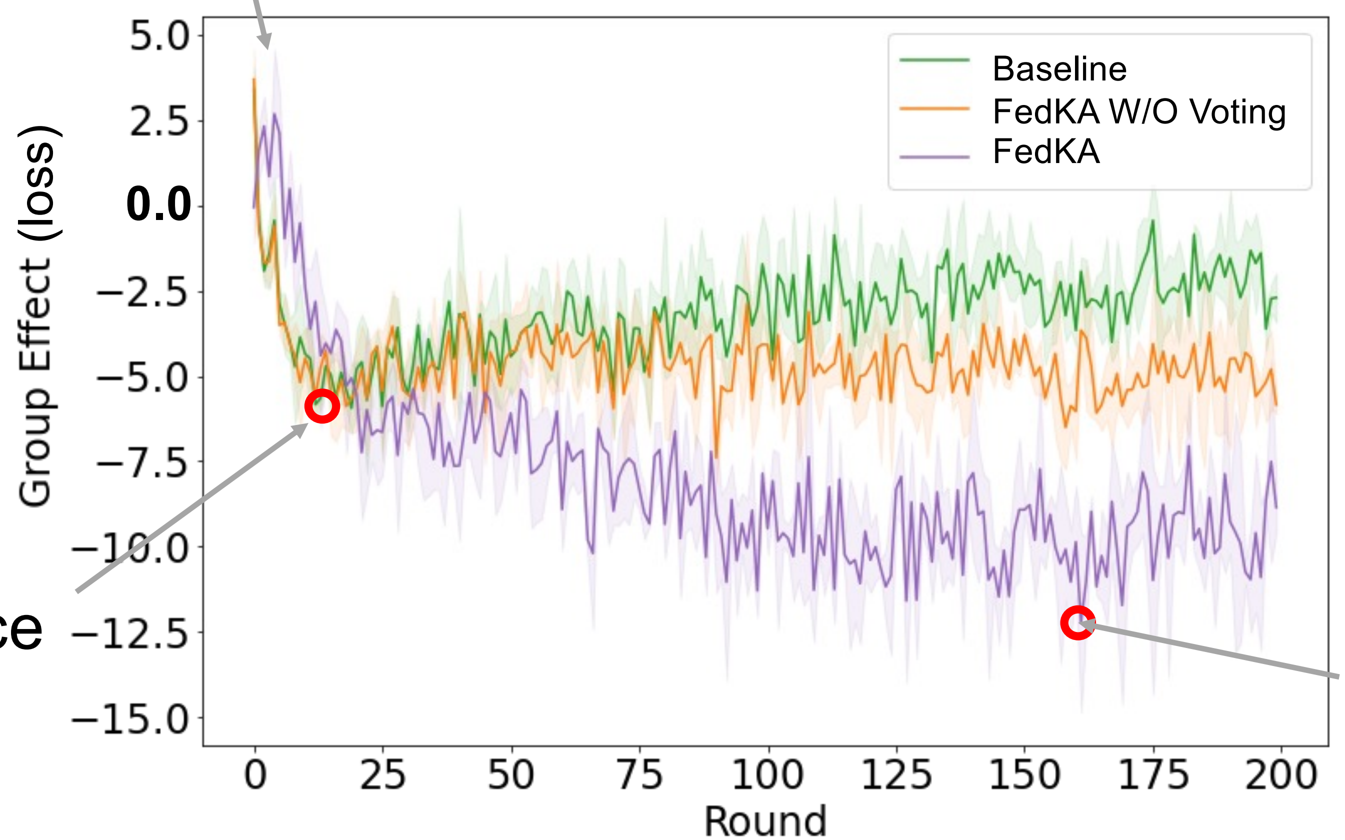
Kitchen: Simple, straight forward to use, very easy to clean, and durable

Late convergence

Group Effect:

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

Initialized, high **loss**



Convergence

Late convergence

Performance evaluation

Digit Five

+4.0%

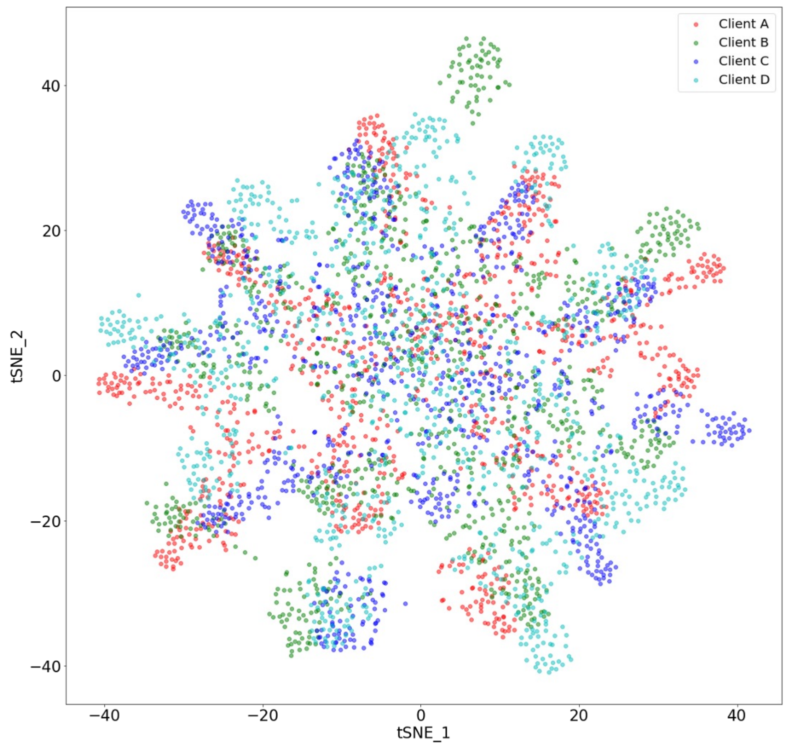
Models/Tasks	→mt	→mm	→up	→sv	→sy	Avg
FedAvg	93.5±0.15	62.5±0.72	90.2±0.37	12.6±0.31	40.9±0.50	59.9
f-DANN	89.7±0.23	70.4±0.69	88.0±0.23	11.9±0.50	43.8±1.04	60.8
f-DAN	93.5±0.26	62.1±0.45	90.2±0.13	12.1±0.56	41.5±0.76	59.9
Voting-S	**93.7**±0.18	63.4±0.28	**92.6**±0.25	14.2±0.99	45.3±0.34	61.8
Voting-L	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**

Office-Caltech10

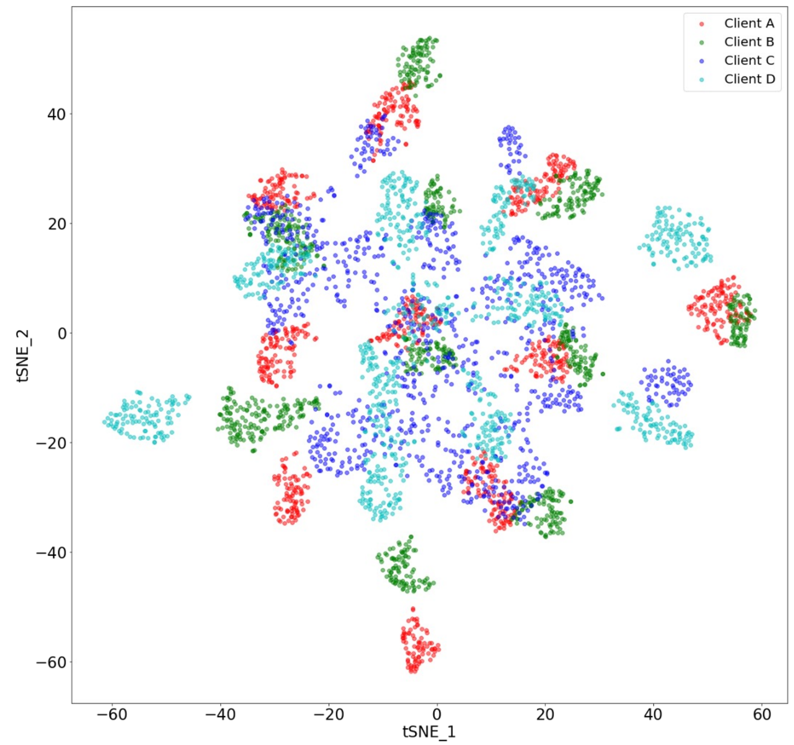
+2.3%

Models/Tasks	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

T-SNE visualization of representation distributions

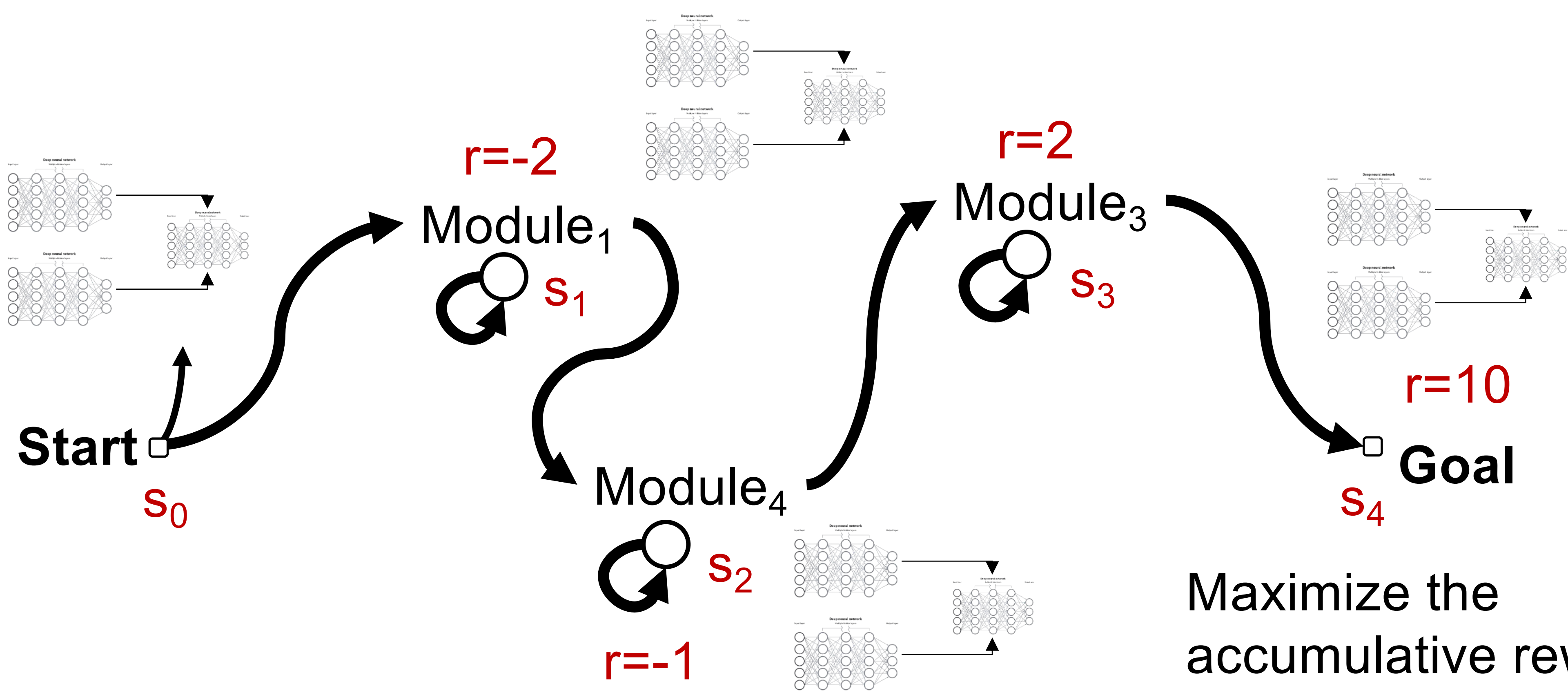


Without knowledge transfer



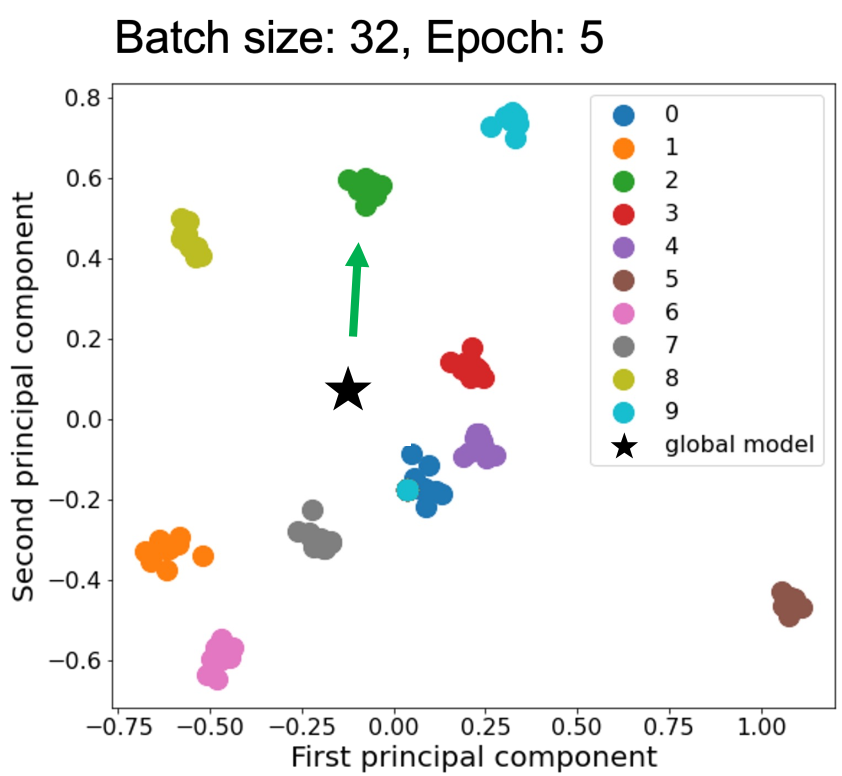
With knowledge transfer

Hierarchical learning as a Markov decision process



Maximize the accumulative reward

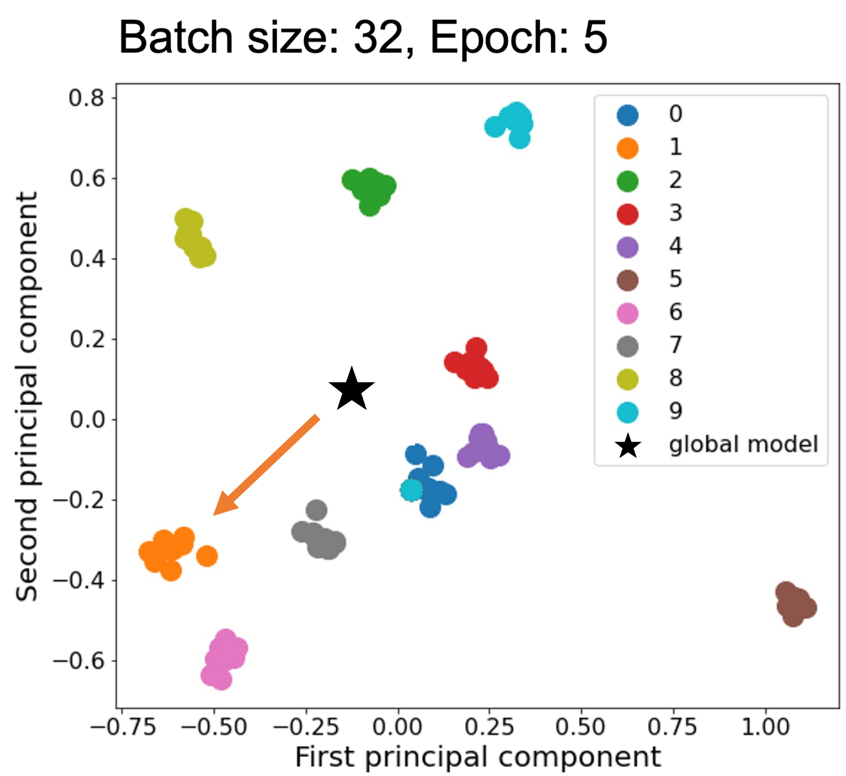
Neural states of modules



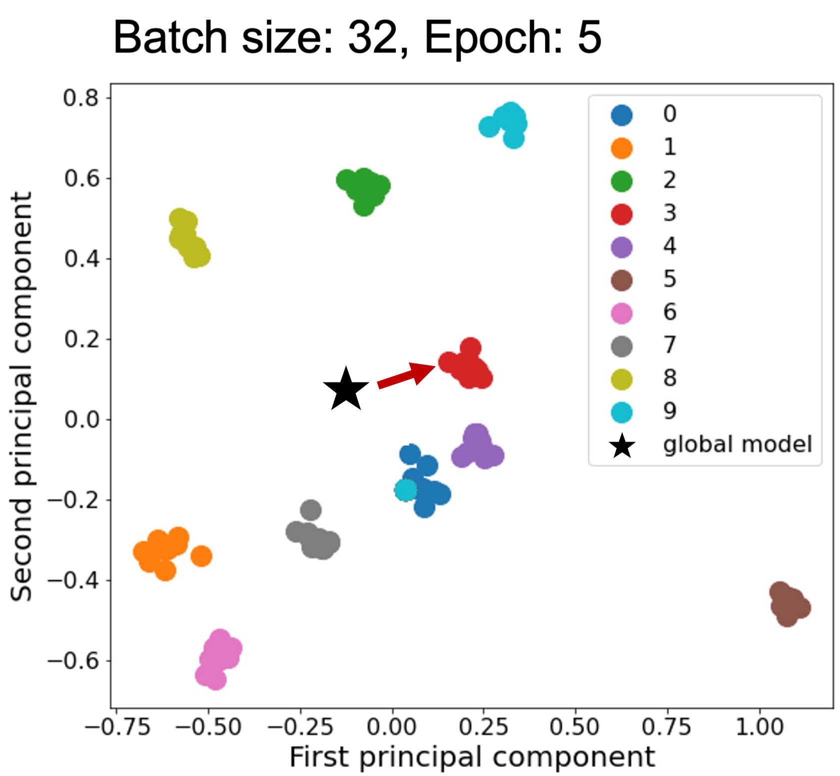
- Problem: Non-independent and identically distributed data (non-IID)
- Cluster: Compressed weights of modules with similar training data classes
- Reliable incremental measures of progress

Yuwei Sun, Hideya Ochiai. Homogeneous Learning: Self-Attention Decentralized Deep Learning. IEEE Access. 2022.

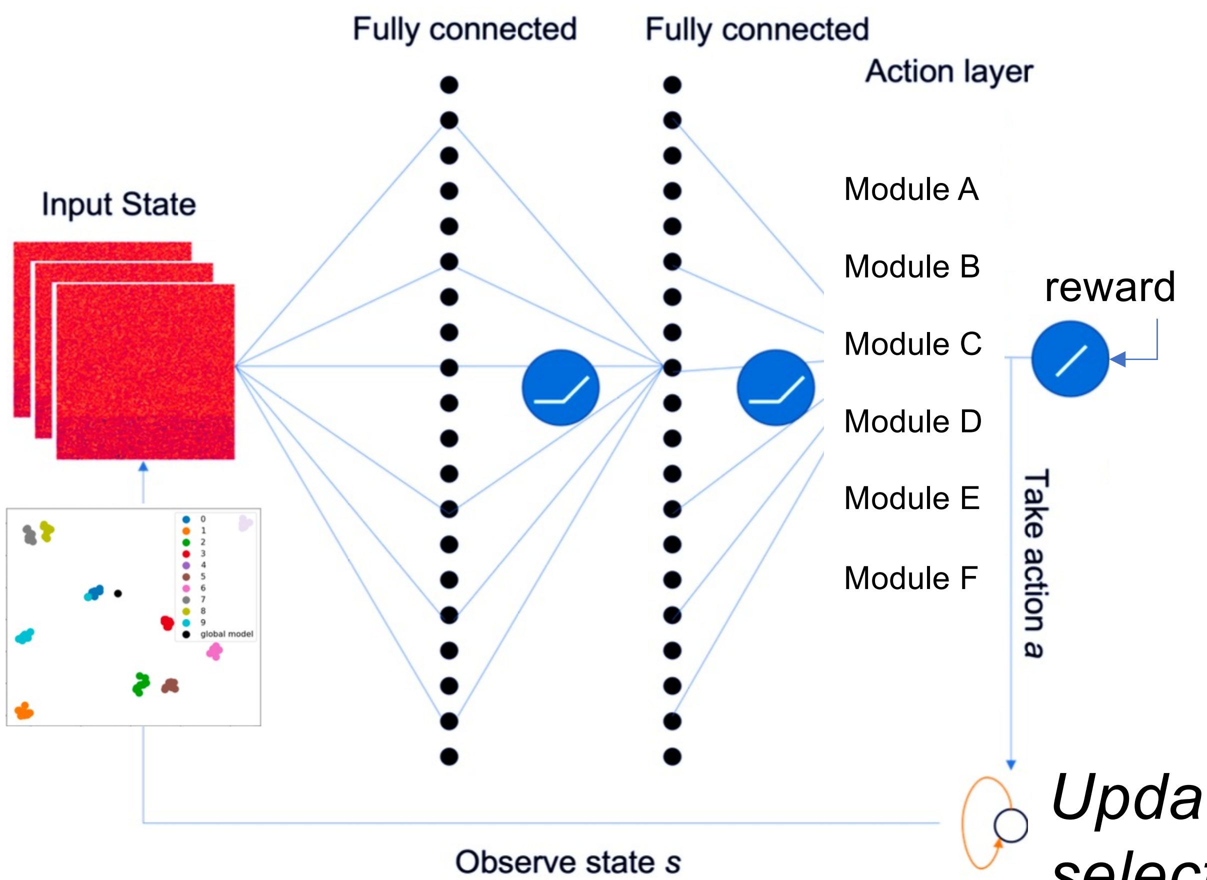
Neural states of modules



Neural states of modules



Outer loop reinforcement learning



States Previous action

$$S_t = f_{HL}(S_{t-1}, \hat{a}_{t-1})$$

Policy Reward

$$\widehat{\theta}_t = \arg\max_{\theta_t}(r_t(S_t) + \gamma \cdot \hat{r}_{t+1}(S_{t+1}))$$

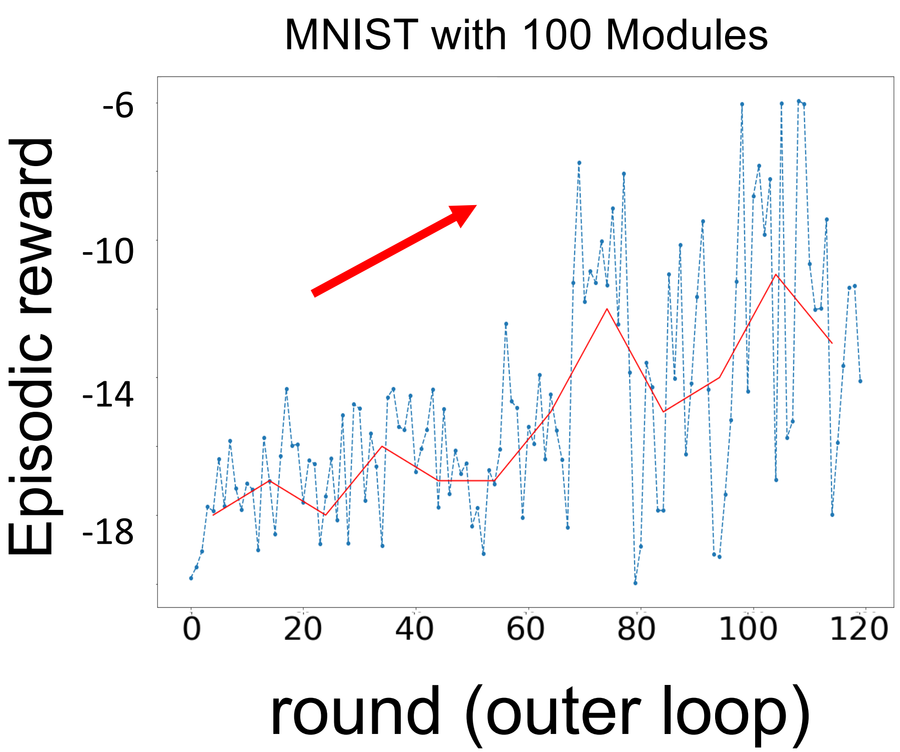
Next action

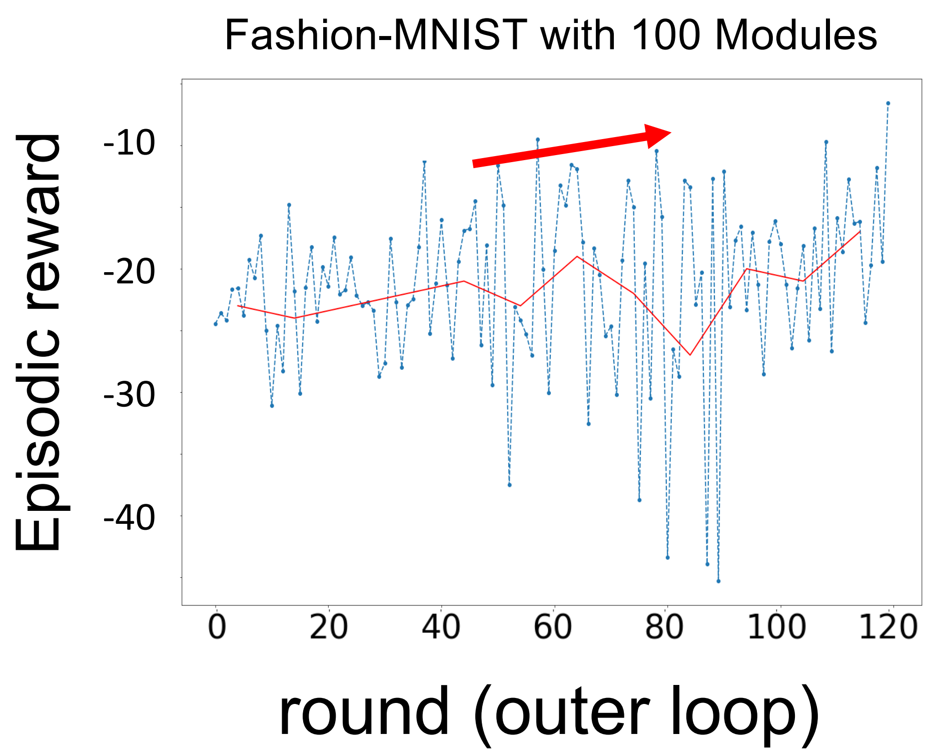
$$\widehat{a}_t = \arg\max_{a_t}(f_{RL}(S_t, \widehat{\theta}_t))$$

Update using the selected module

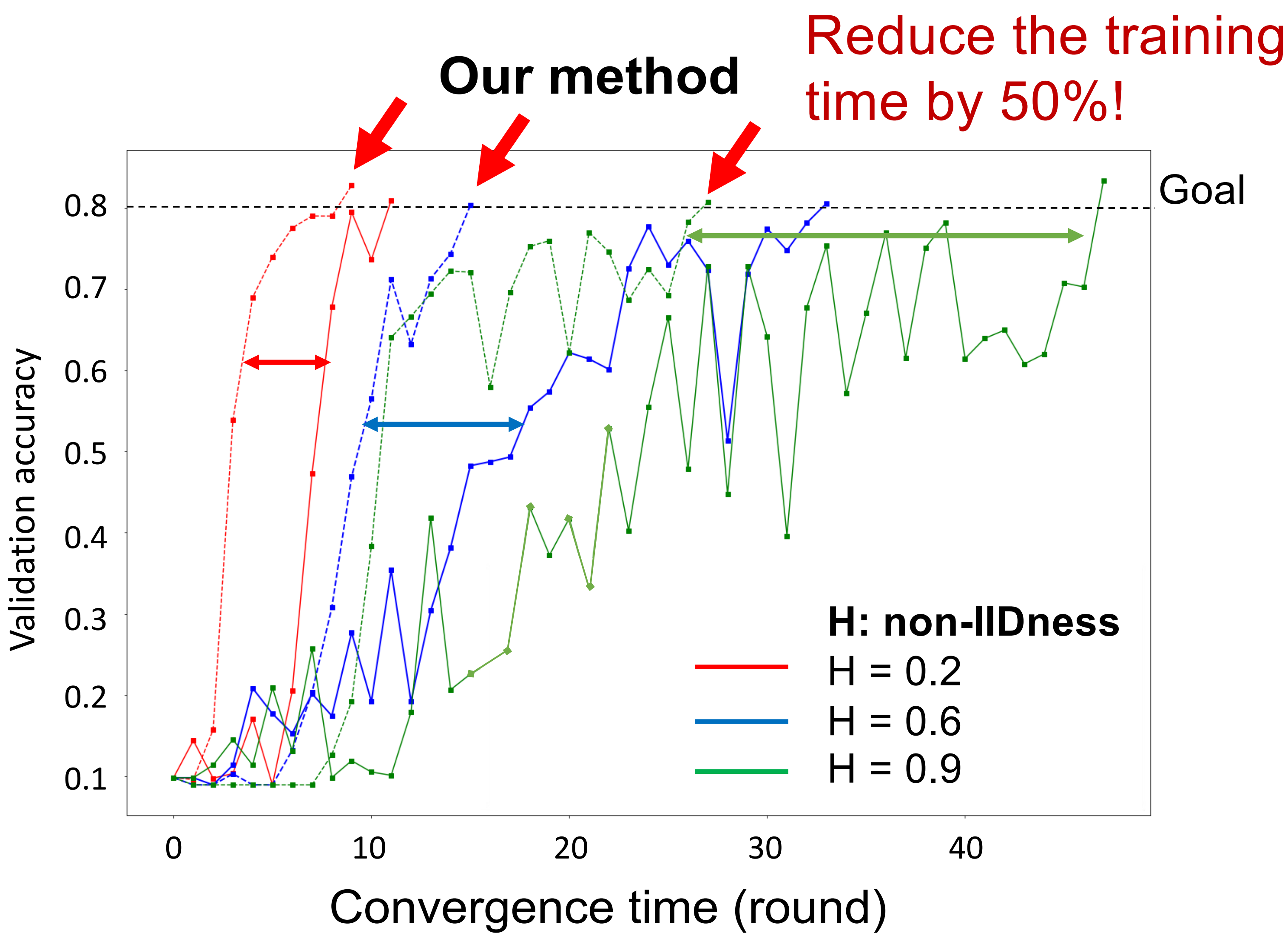
Yuwei Sun, Hideya Ochiai. Homogeneous Learning: Self-Attention Decentralized Deep Learning. IEEE Access. 2022.

Reward learning





Reduced convergence time



Cross-modal knowledge transfer

- World model with different modalities
- Cross-modal knowledge transfer
- Visual Question Answering tasks

Question:

What shape are the pizzas?

Answer: square

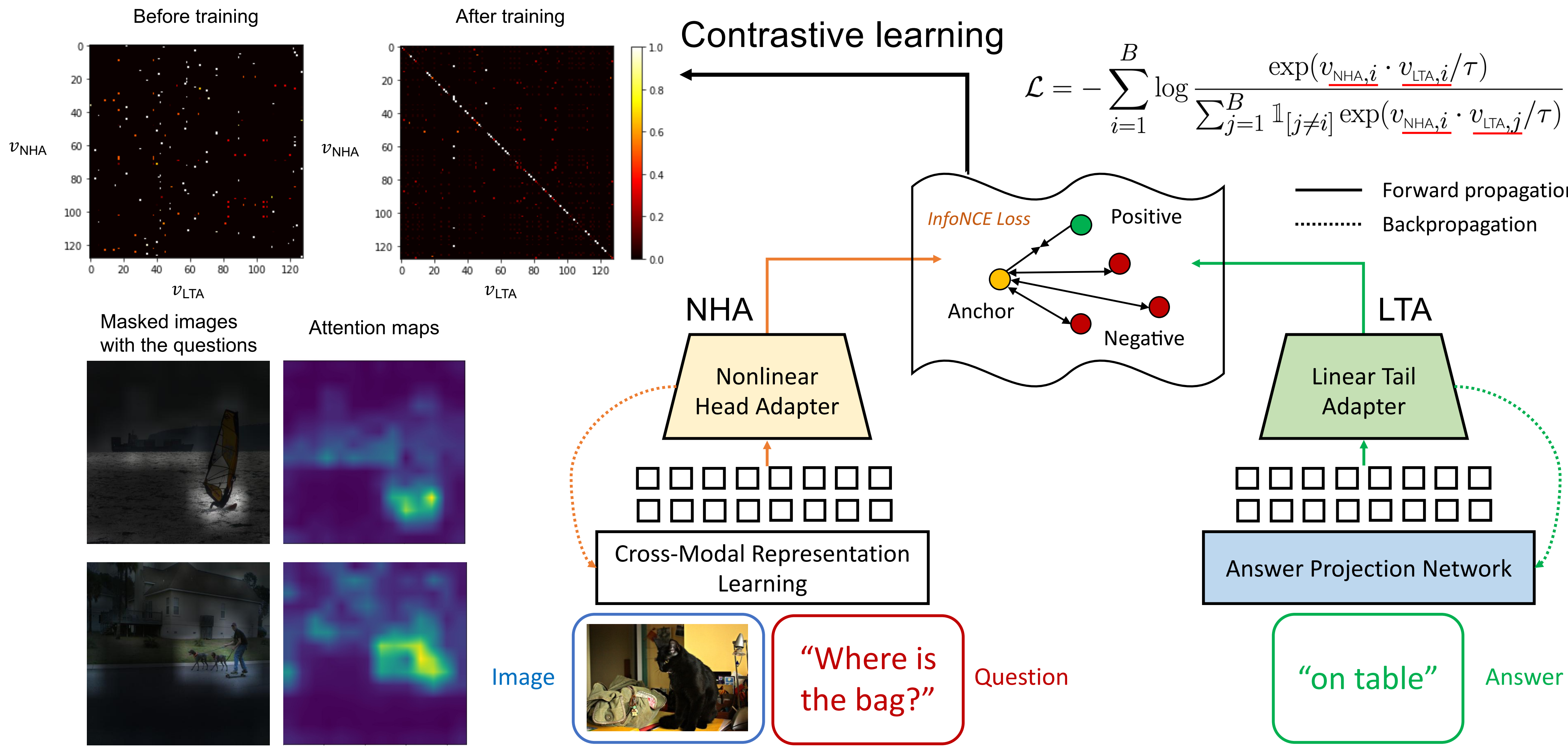
Multi-agent Visual Question Answering

vs.

Transferred knowledge

- Training on the entire distribution
- Subsets representing different perspectives

Modality alignment with self-supervised learning



Yuwei Sun, Hideya Ochiai. UniCon: Unidirectional Split Learning with Contrastive Loss for Visual Question Answering. NeurIPS 2022 workshop

Evaluation

VQA Models	Contrastive learning-based VQA (%)			
	Overall	Yes/No	Number	Other
BAN	36.23	66.90	12.71	19.11
BUTD	45.08	75.82	29.27	25.86
MFB	46.98	73.95	32.81	30.20
MCAN-s	53.18	81.06	41.95	34.93
MCAN-l	53.32	**81.21**	42.66	34.90
MMNas-s	51.54	78.06	39.76	34.46
MMNas-l	$\boxed{\textbf{53.82}}$	80.06	**42.86**	**36.75**

VQA Models	UniCon (%)			
	Overall	Yes/No	Number	Other
BAN	35.11	63.84	11.06	19.61
BUTD	40.96	66.98	13.34	28.74
MFB	42.43	68.65	23.33	27.52
MCAN-s	48.42	74.93	30.88	32.89
MCAN-l	48.44	**77.44**	30.72	32.01
MMNas-s	45.14	70.55	28.04	30.33
MMNas-l	$\boxed{\textbf{49.89}}$	74.85	**36.88**	**34.33**

Q: Which room is this?
A: bedroom
Ground Truth: bedroom

Q: How many pictures on the wall?
A: 6
Ground Truth: 7

Conclusions

- Limitation: Out-of-distribution generalization ability of NNs
- Benefits of knowledge sharing and social learning among NN models in unseen tasks
- Network of interconnected NN models with similar architecture
- Hierarchical NNs with a meta model to optimize policy
- Self-supervised learning for cross-modal knowledge transfer without labels

Reusable modular knowledge for systematic generalization

- Decompose high-level knowledge into **reusable components**
- Attention mechanism
- Switch from System 1 to System 2 processing
- **Routing** of reusable components to tackle the OOD problem
- Graph-structured elements of **causality**
- Interventions, effects, and **interpretability**

Meta Learning in Decentralized Neural Networks Towards More General AI

Yuwei Sun

The University of Tokyo

RIKEN

東京大学
THE UNIVERSITY OF TOKYO

AIP
Center for
Advanced Intelligence Project