Meta Neural Coordination in Decentralized Neural Networks

Yuwei Sun

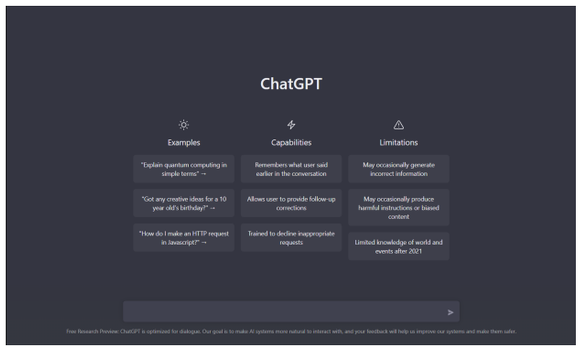
The University of Tokyo

RIKEN AIP

The state of deep learning

Go

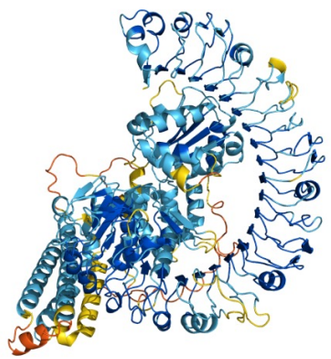
Large language model



Self-driving car

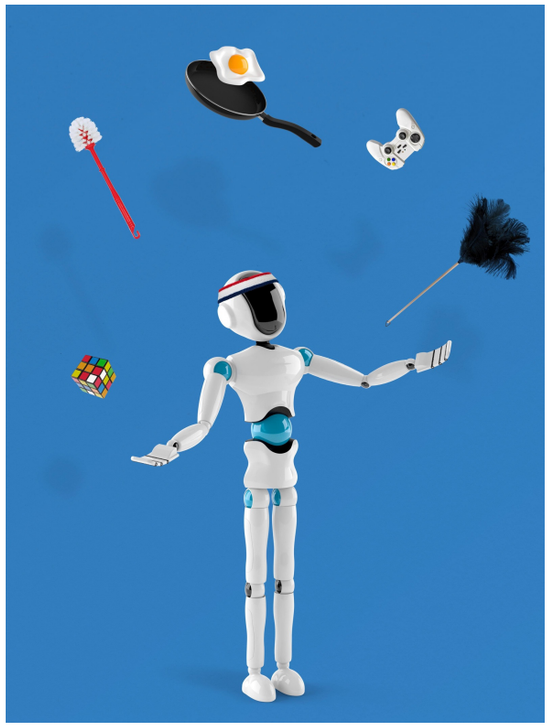
Text to video

Protein folding



Out-of-distribution generalization

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

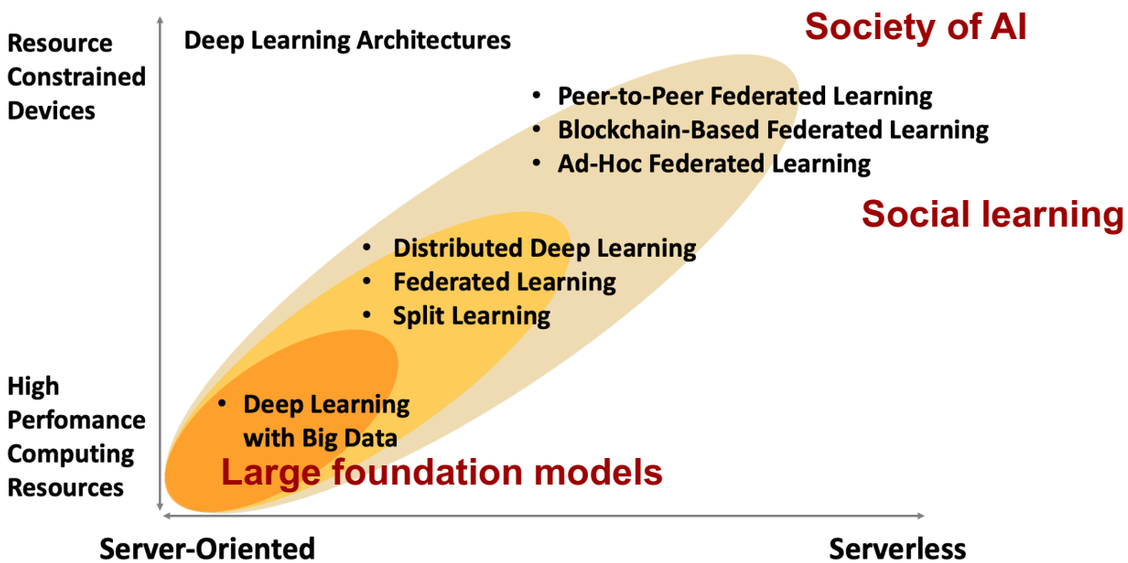


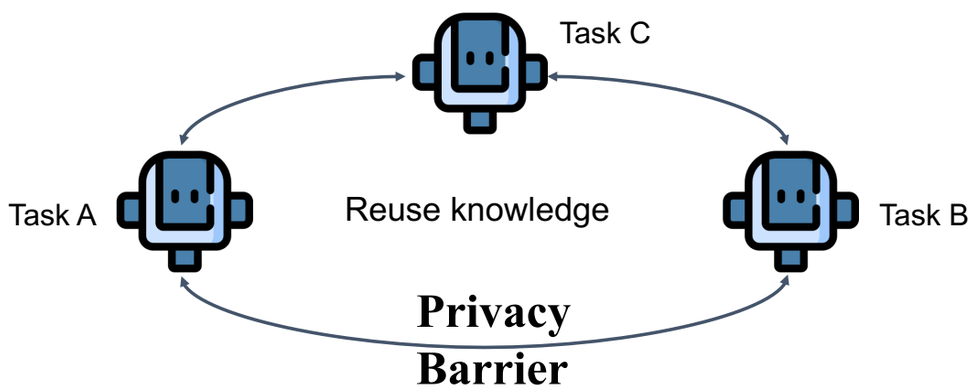
ACM

Beyond the training distribution

Why decentralized ML?

- 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

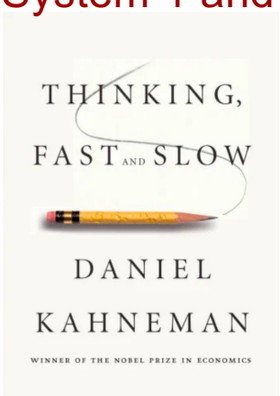


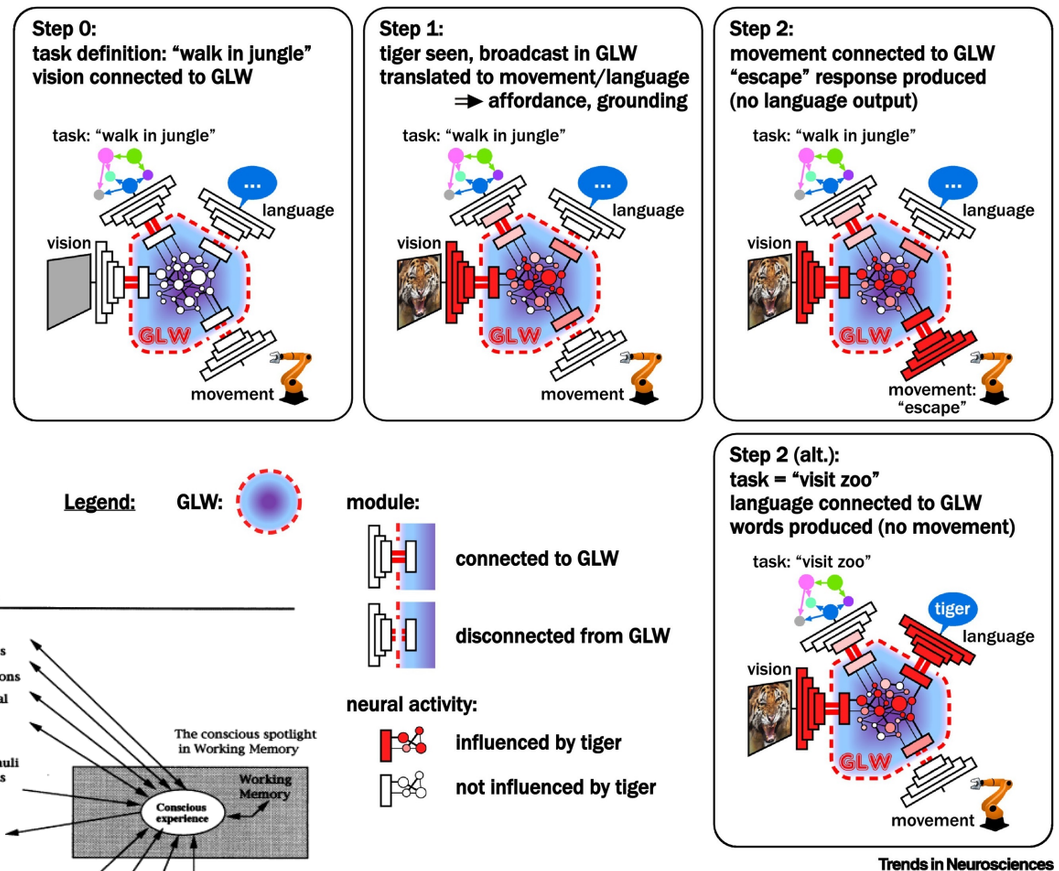


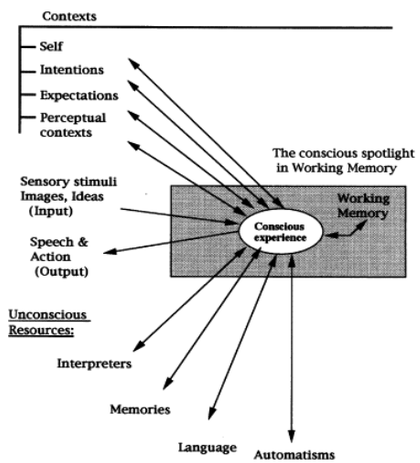
Global Workspace Theory

- Cooperating and competing neural network models
- Specialized processors
- Selection and reuse of processors

System 1 and System 2 AI



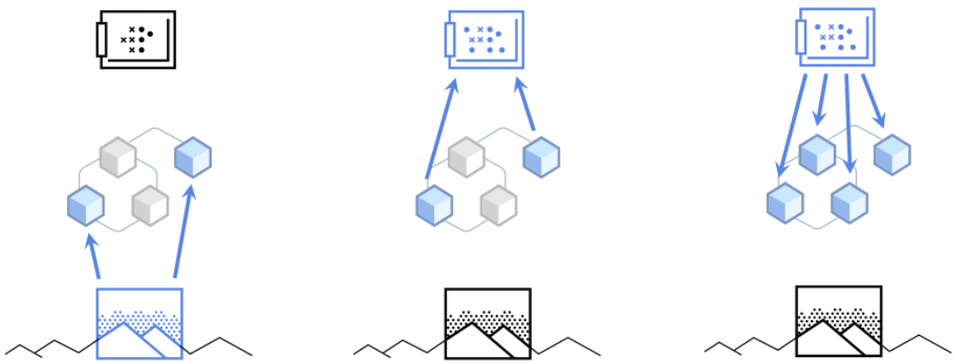




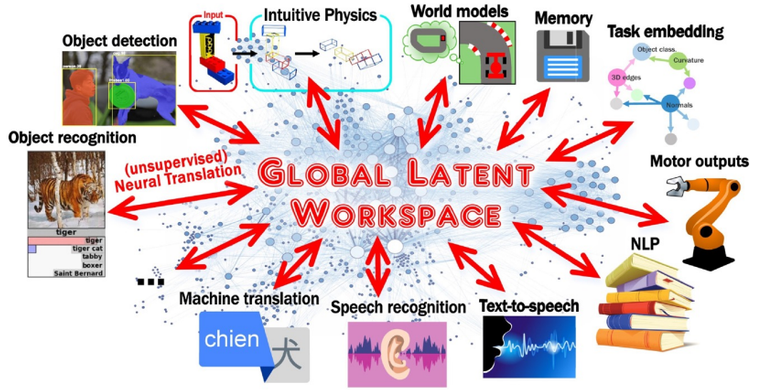
[Baars, 1988]

[Kanai, 2021]

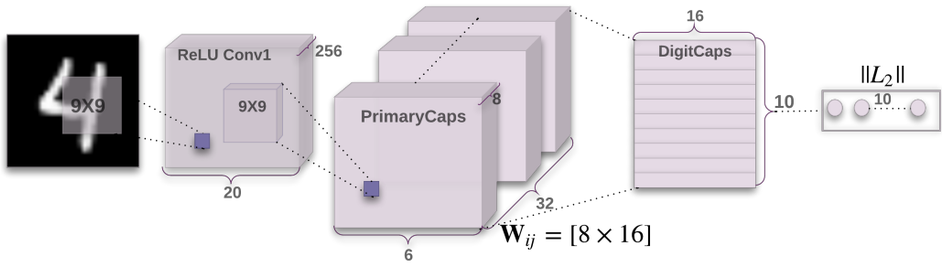
Coordination in Neural Modules



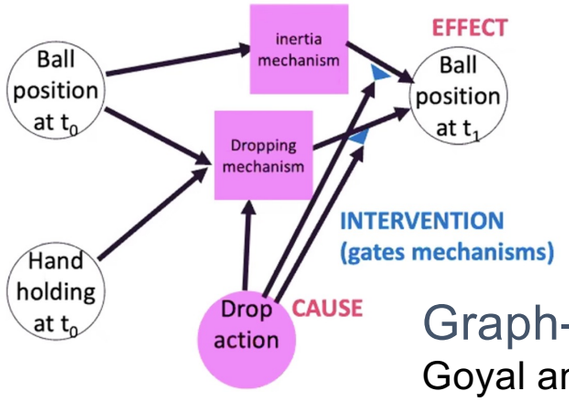
Global Workspace Theory
Bengio et al., ICLR'22



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



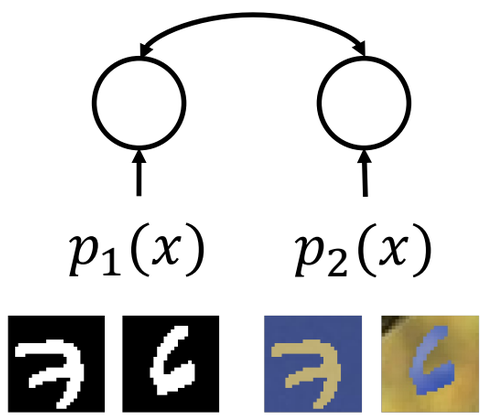
Capsule Networks
Hinton et al., NeurIPS'17



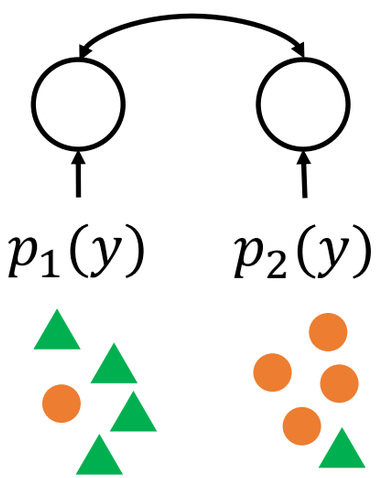
Graph-structured causality
Goyal and Bengio, arXiv'22

Reusable knowledge representation learning

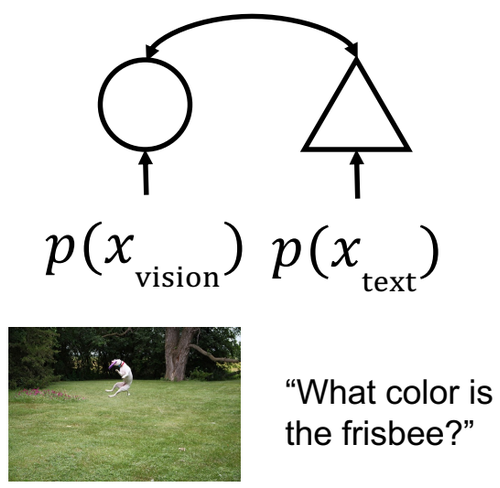
- Enhancing generalization through module knowledge transfer



(a) distributional shift



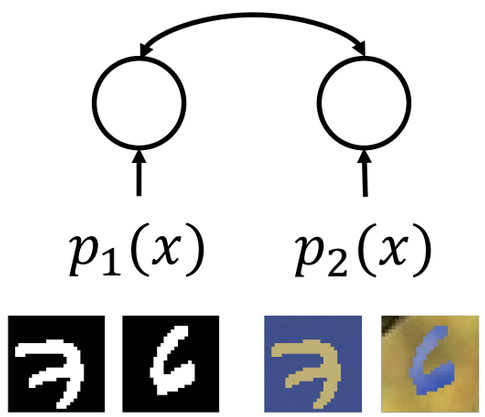
(b) class shift



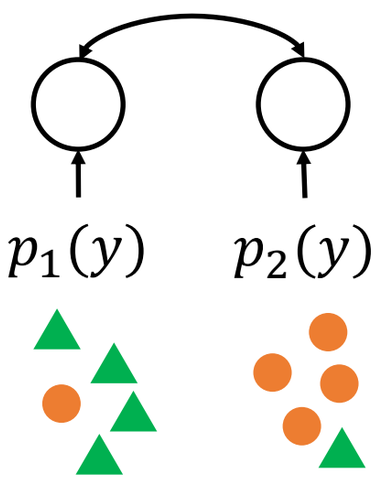
(c) cross-modal

Reusable knowledge representation learning

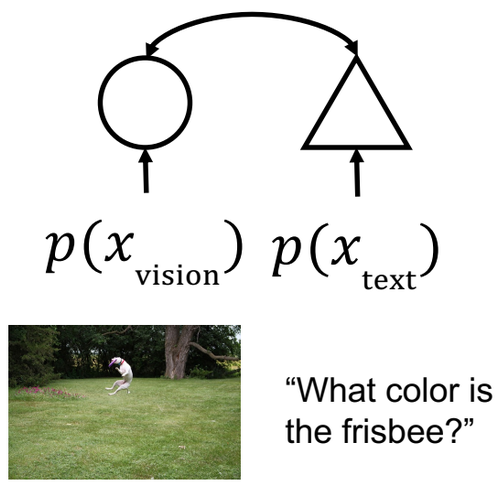
- Enhancing generalization through module knowledge transfer



(a) distributional shift



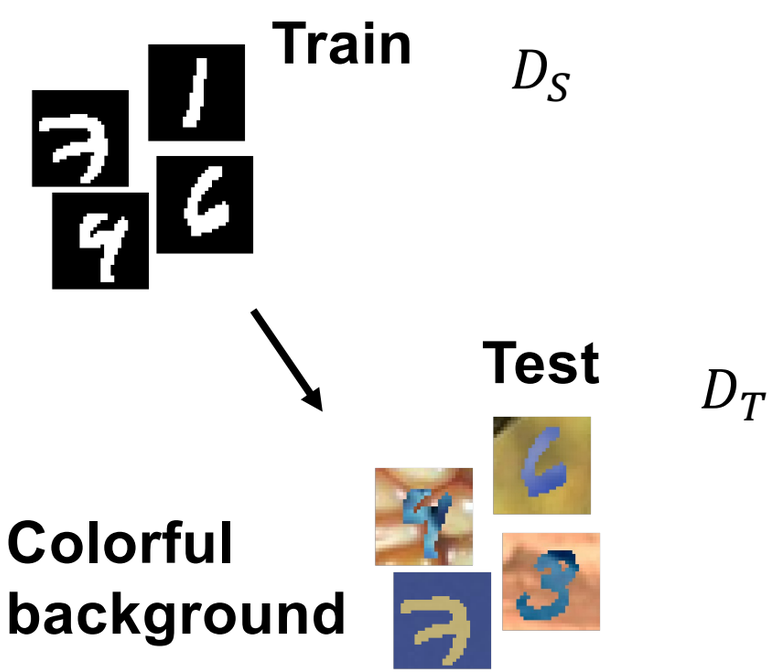
(b) class shift

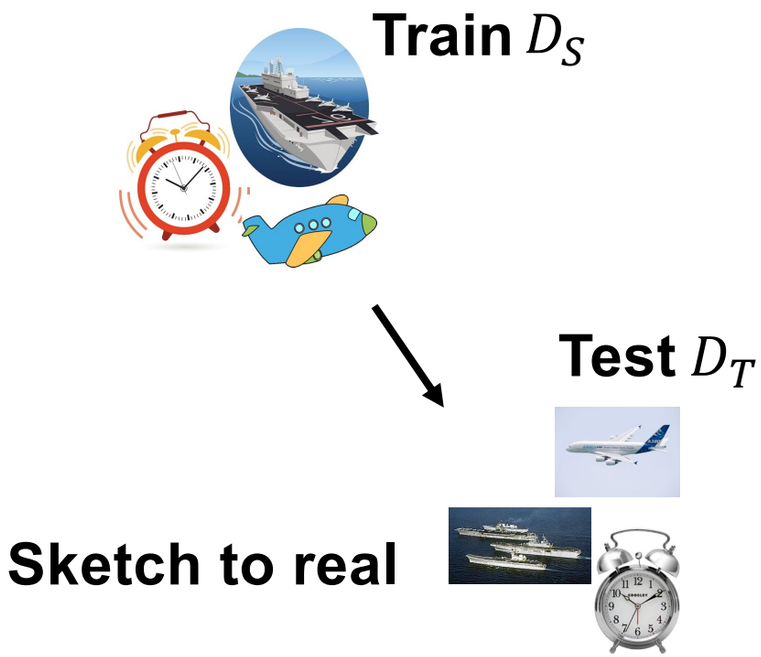


(c) cross-modal

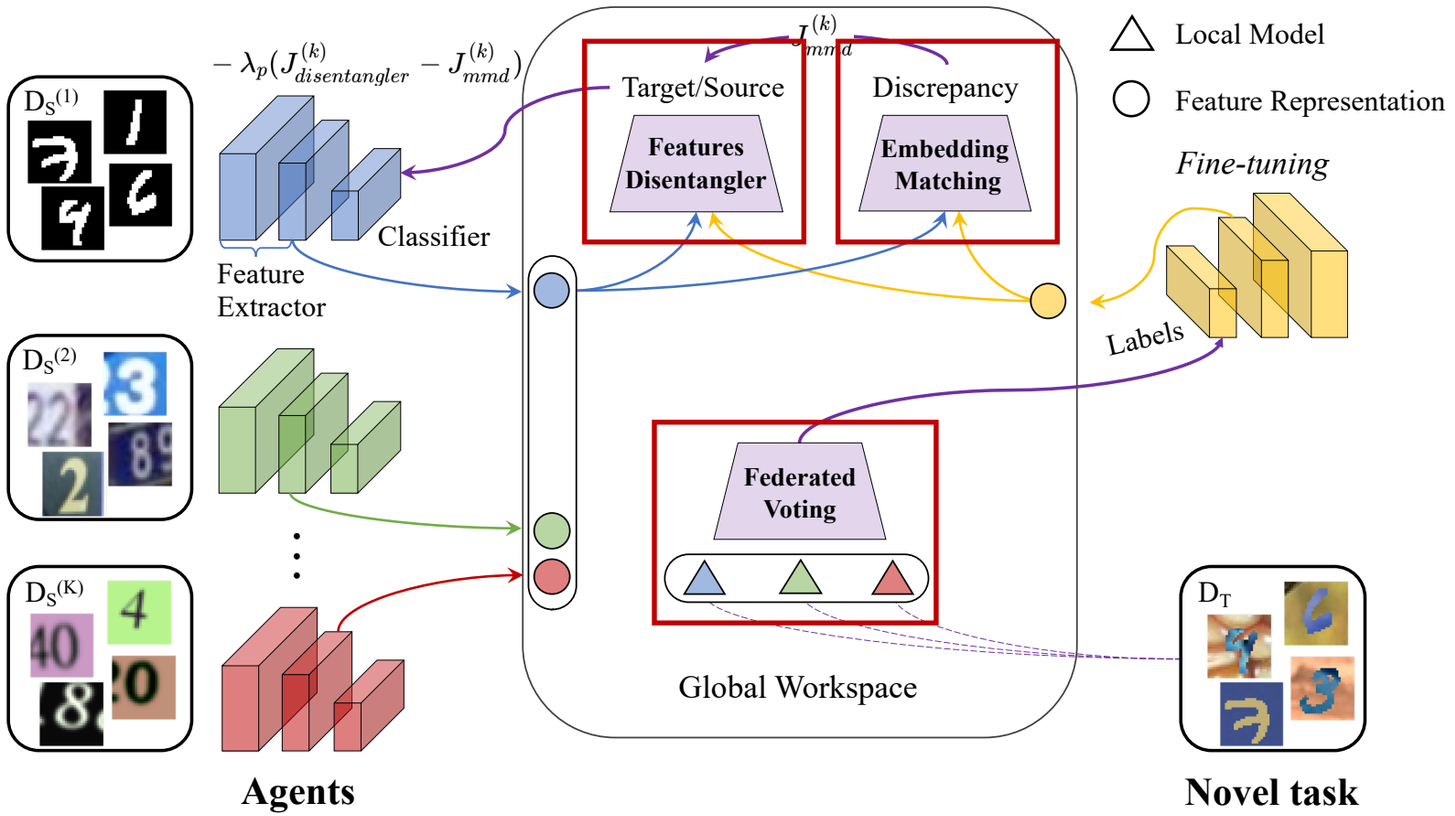
Distributional shift between train and test dataset

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



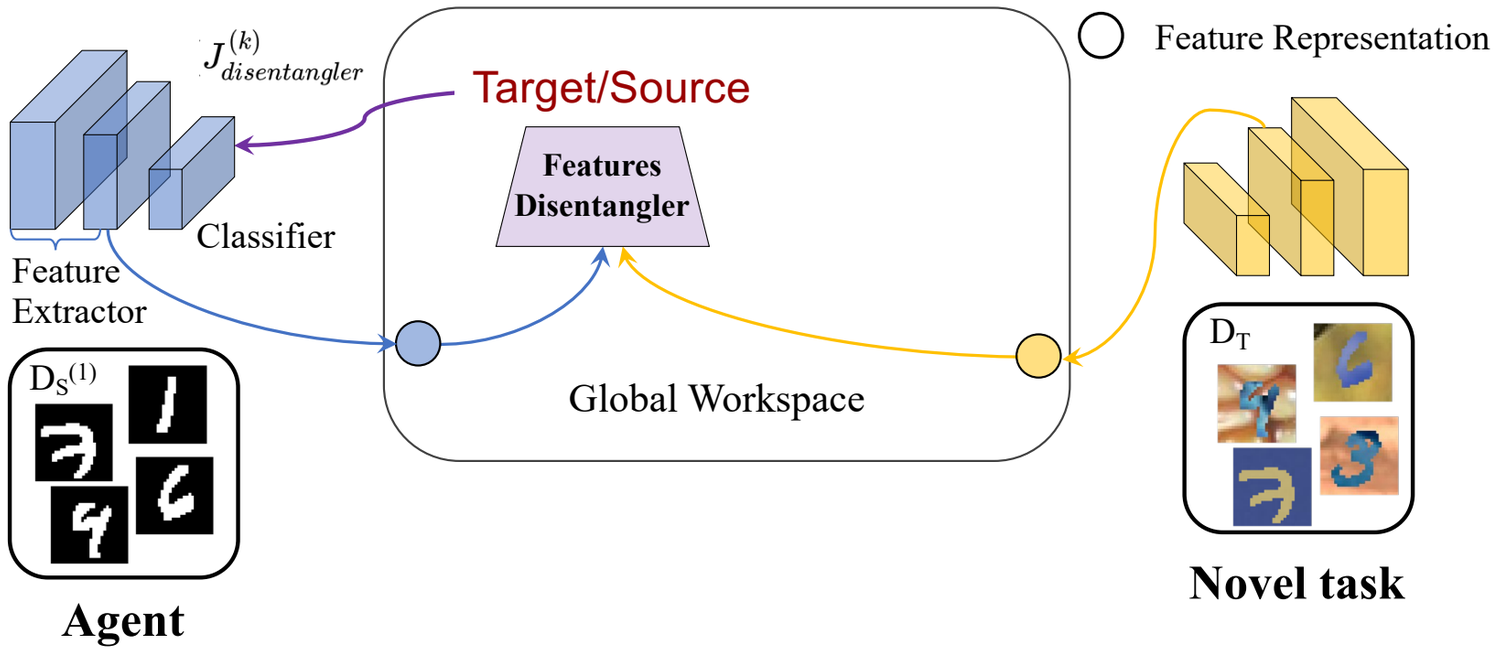


Feature distribution matching for federated domain generalization



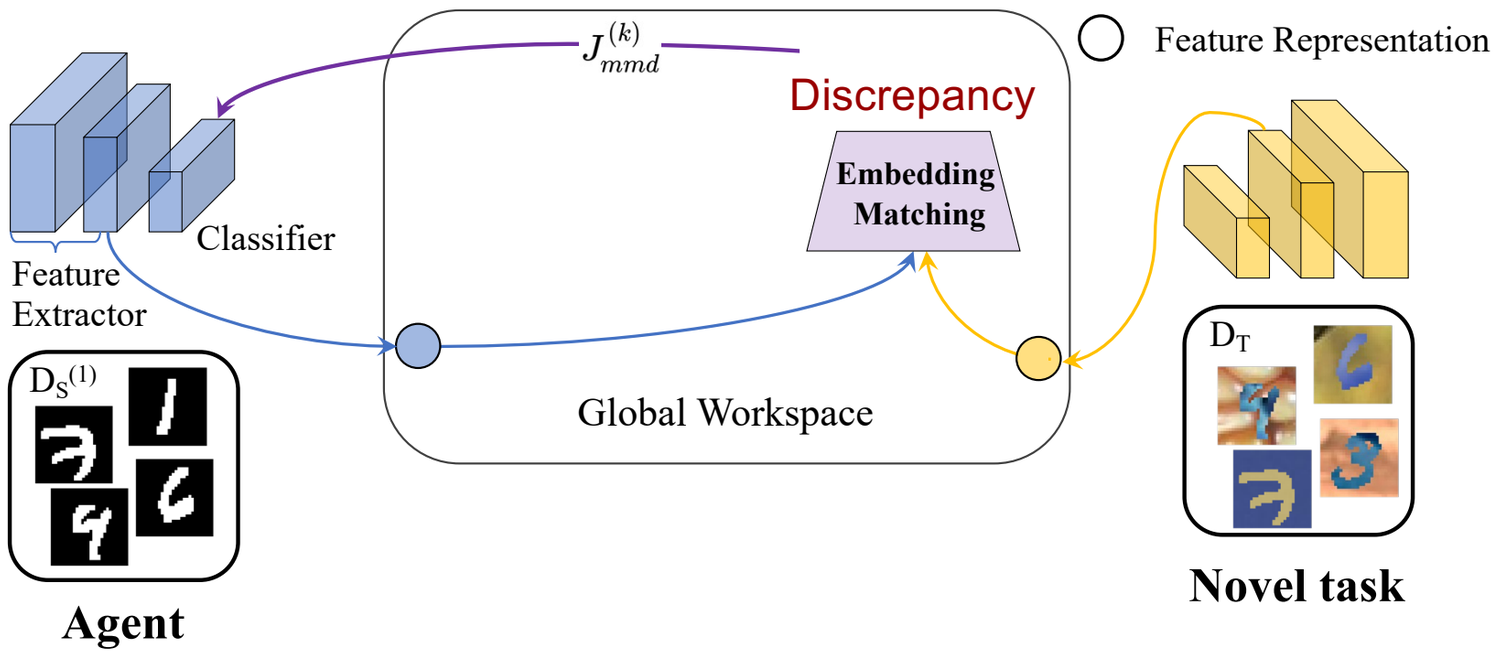
Sun et al. Feature Distribution Matching for Federated Domain Generalization, ACML 2022.

Global features disentangler



- f_d distinguishes between domain representations of the source and the target f_e^k and f_e^G $\quad \hat{f}_e^k = \arg\max_{f_e^k} J^{(k)}_{\text{disentangler}}(\hat{f}_d, f_e^k, \hat{f}_e^G)$
- f_e^k learns to extract **domain-invariant** features

Embedding matching



- Multi-kernel maximum mean discrepancy between f_e^k and f_e^G
- f_e^k learns **stable** common features

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

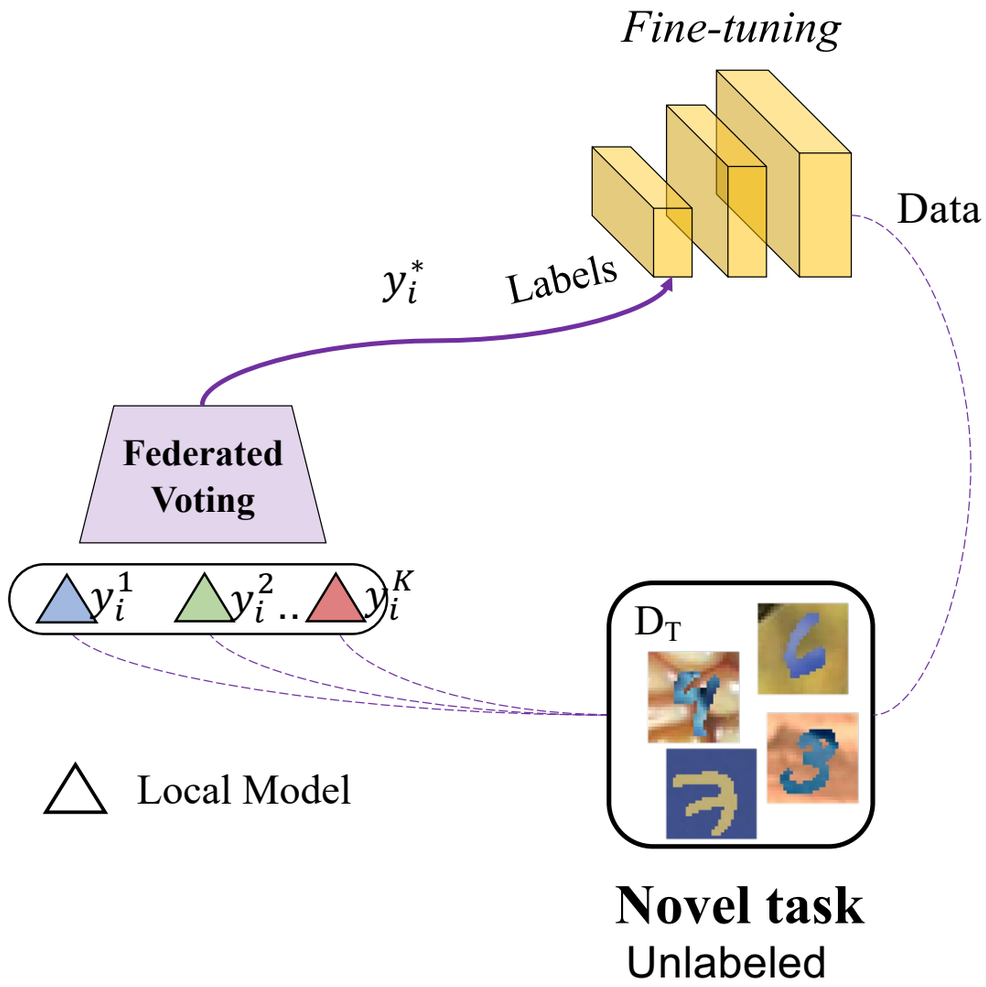
$$\hat{f}_e^k = \arg\min_{f_e^k} j_{\text{mmd}}^k(f_e^k, \hat{f}_e^G)$$

Federated voting

- Generate **pseudo-labels** for the novel task using plurality voting

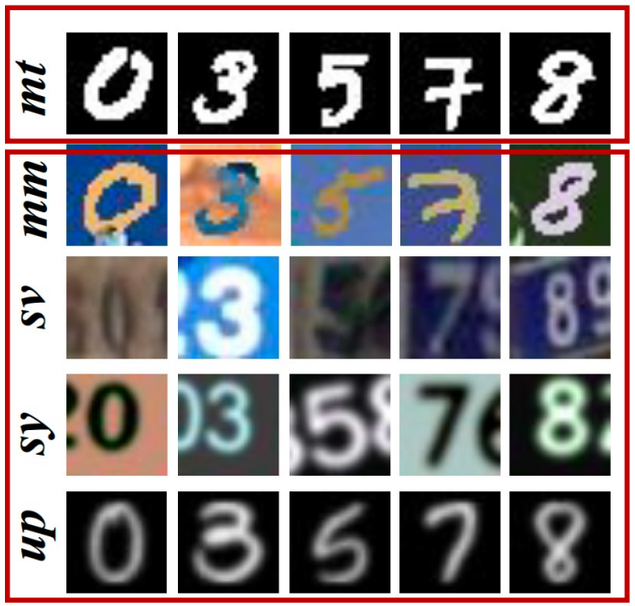
$$y_i^* = \underset{c \in \{1,2,\ldots,C\}}{\arg\max} \sum_{k=1}^{K} \mathbb{I}\{y_i^k = c\}$$

- Fine-tune the global model using data from the new task and the pseudo-labels



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

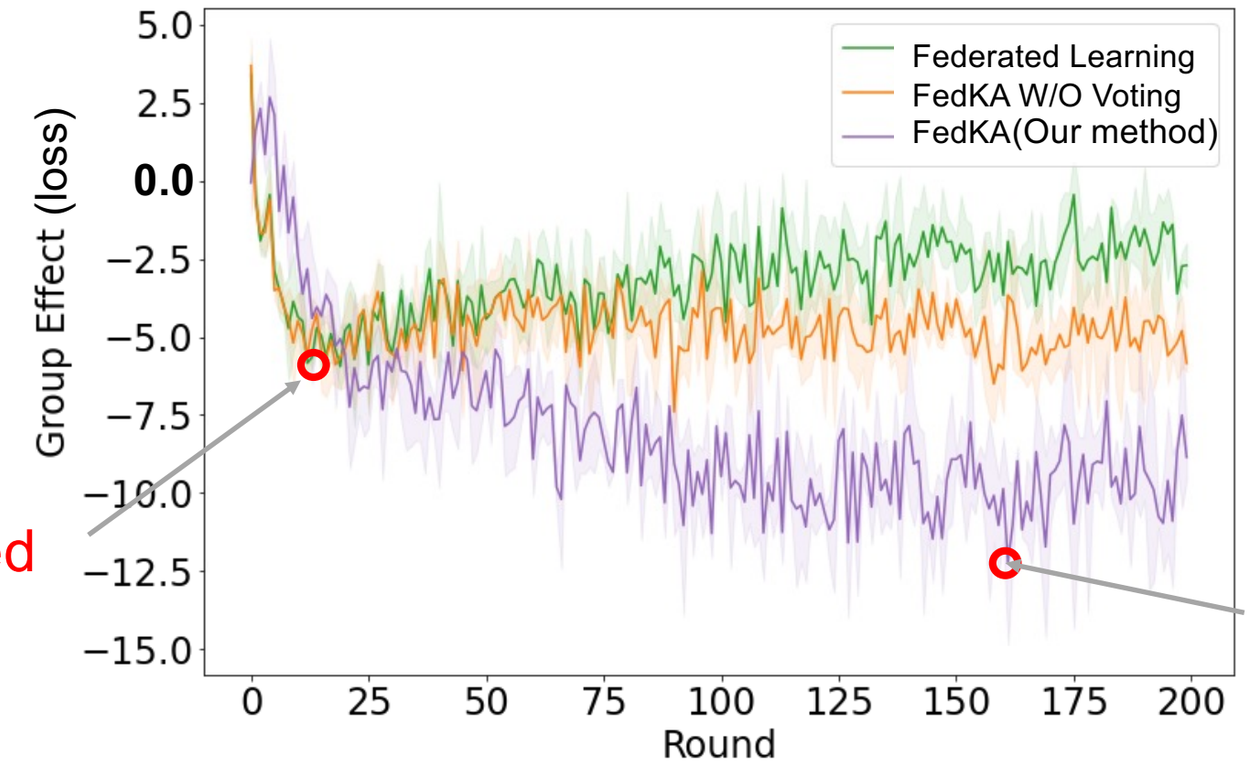
Electronics: This is perfect for my iPod and keeps it totally secure while driving

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

Late convergence

Group Effect (the lower the better):

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



Converged

Late convergence

How much benefit can be derived from knowledge transfer?

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**

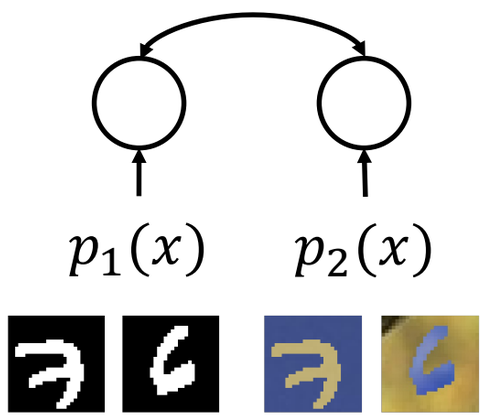
Office-Caltech10

+2.3%

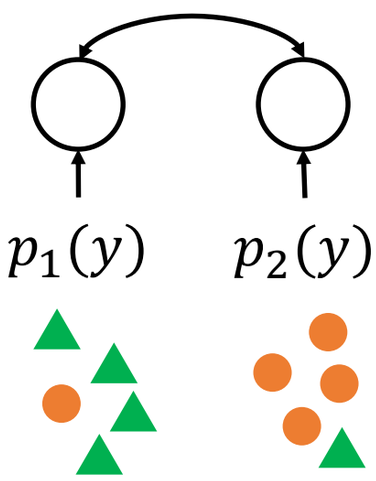
Models/Tasks	C,D,W→A	A,D,W→C	C,A,W→D	C,D,A→W	Avg
FedAvg	56.4 ±1.23	40.2 ±0.69	28.7±1.21	22.7±1.85	37.0
f-DANN	58.3 ±1.53	40.0 ±1.50	30.7 ±3.59	22.3±1.29	37.8
f-DAN	56.7±0.71	38.7±0.75	30.2±1.64	23.9 ±1.70	37.4
Voting	56.5 ±1.88	40.2 ±0.58	29.8±1.45	**24.1** ±0.69	37.7
Disentangler + Voting	**61.4** ±2.51	**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

Reusable knowledge representation learning

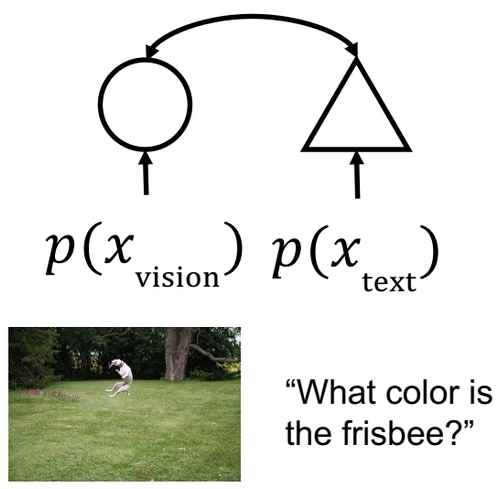
- Enhancing generalization through module knowledge transfer



(a) distributional shift

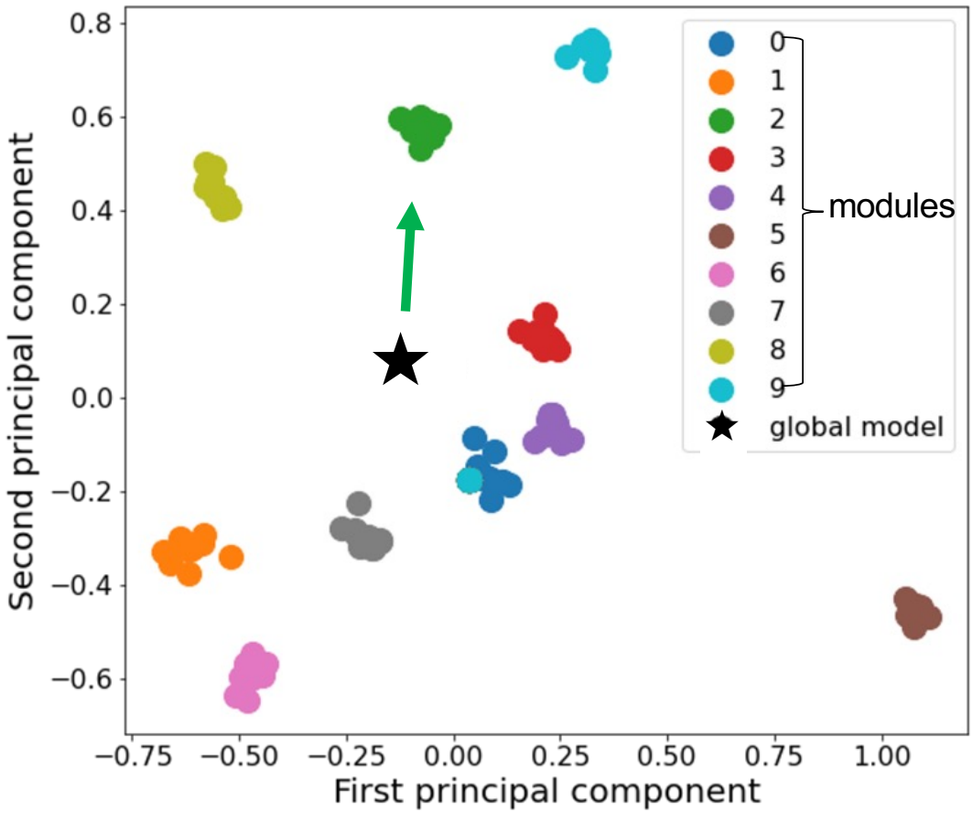


(b) class shift



(c) cross-modal

Non-independent and identically distributed data



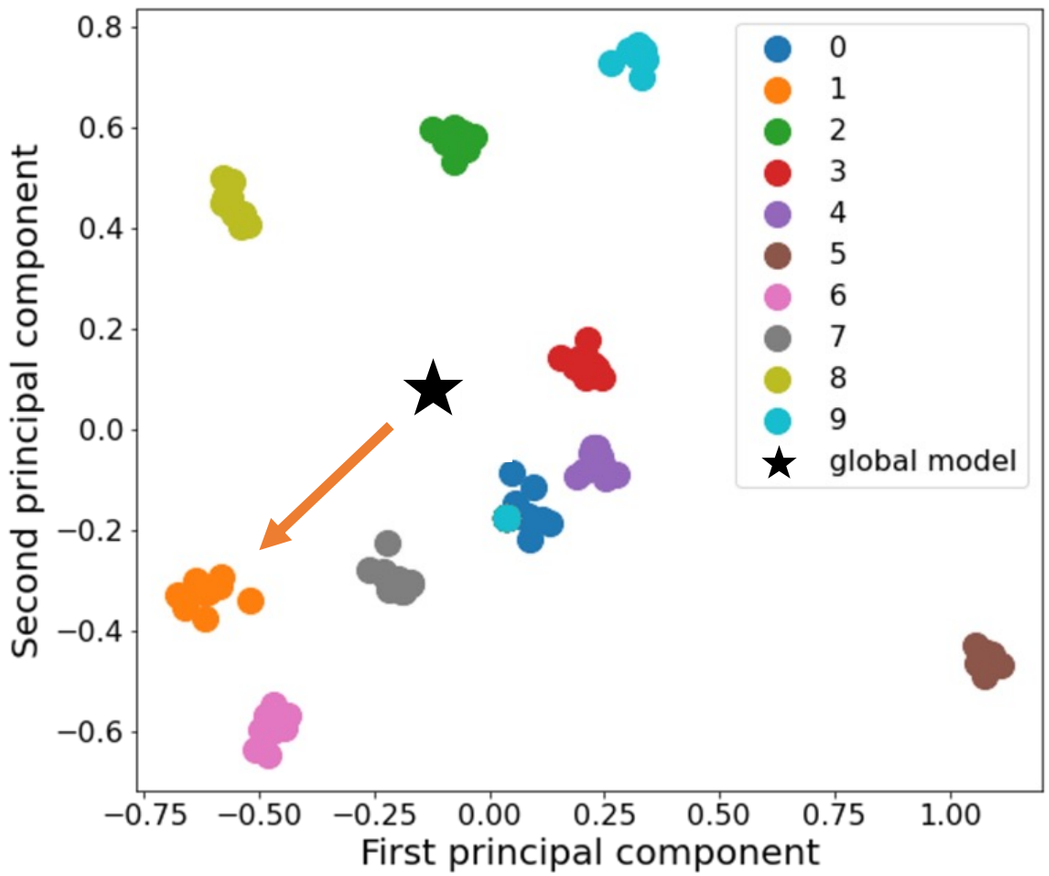
For a classification task involving C classes, each module has its own dataset, with one main class that contains the largest amount of data

$$p_i(x, y) = p(x|y)p_i(y)$$
s.t. $p_i(y) \neq p_j(y) \quad \forall i \neq j,\ p_i\left(y = c^{(i)}\right) > 0.5$

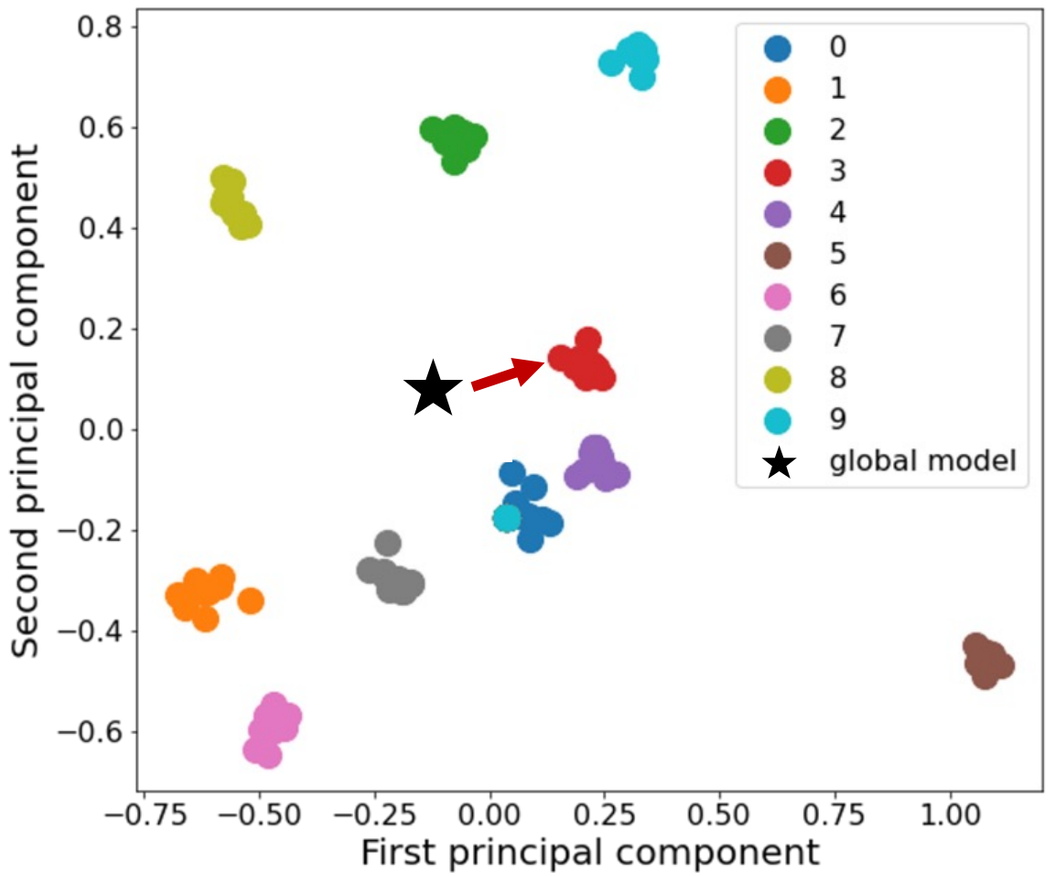
Clusters: Neural network modules trained on similar data class distributions

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

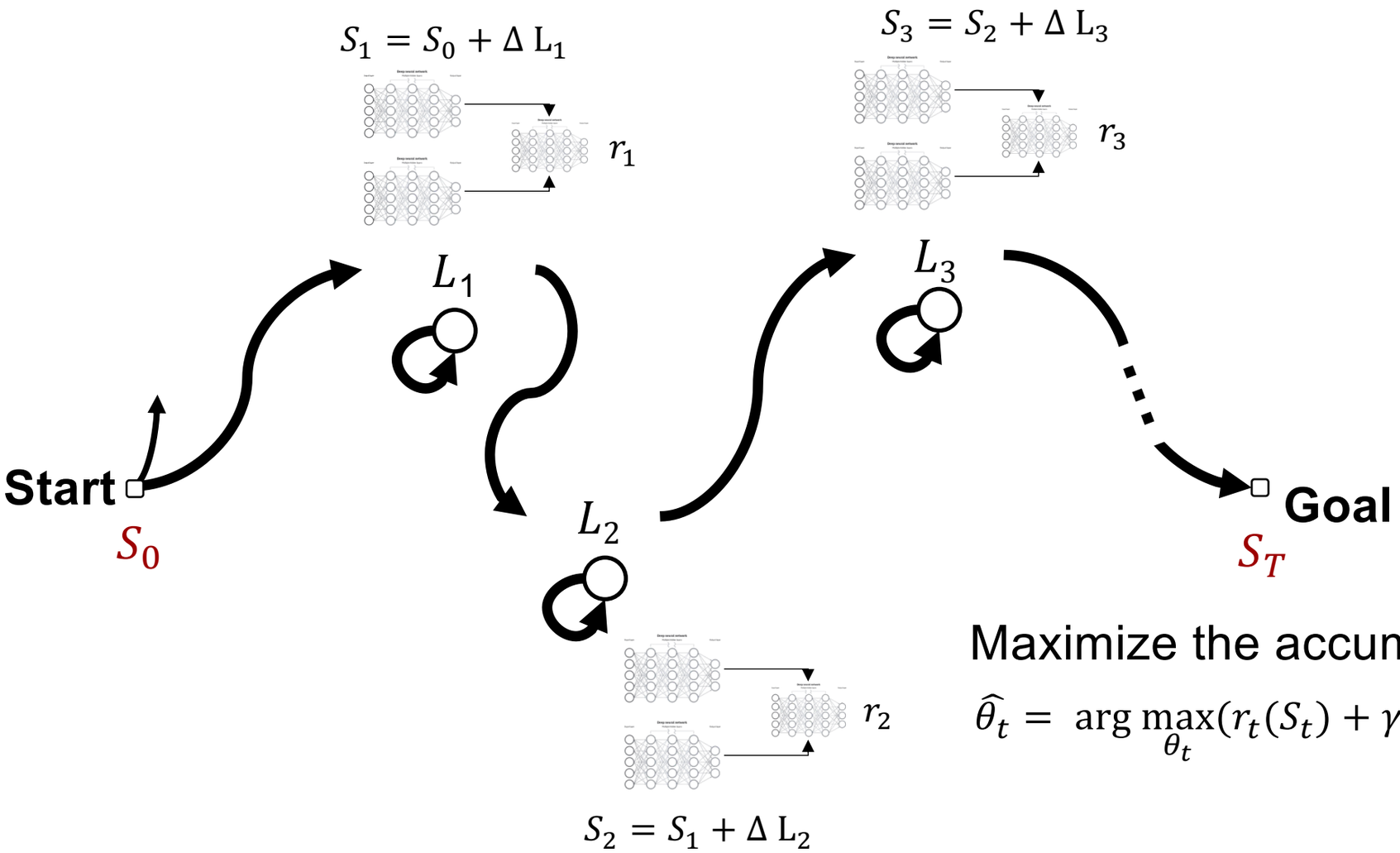
Neural states of modules



Neural states of modules



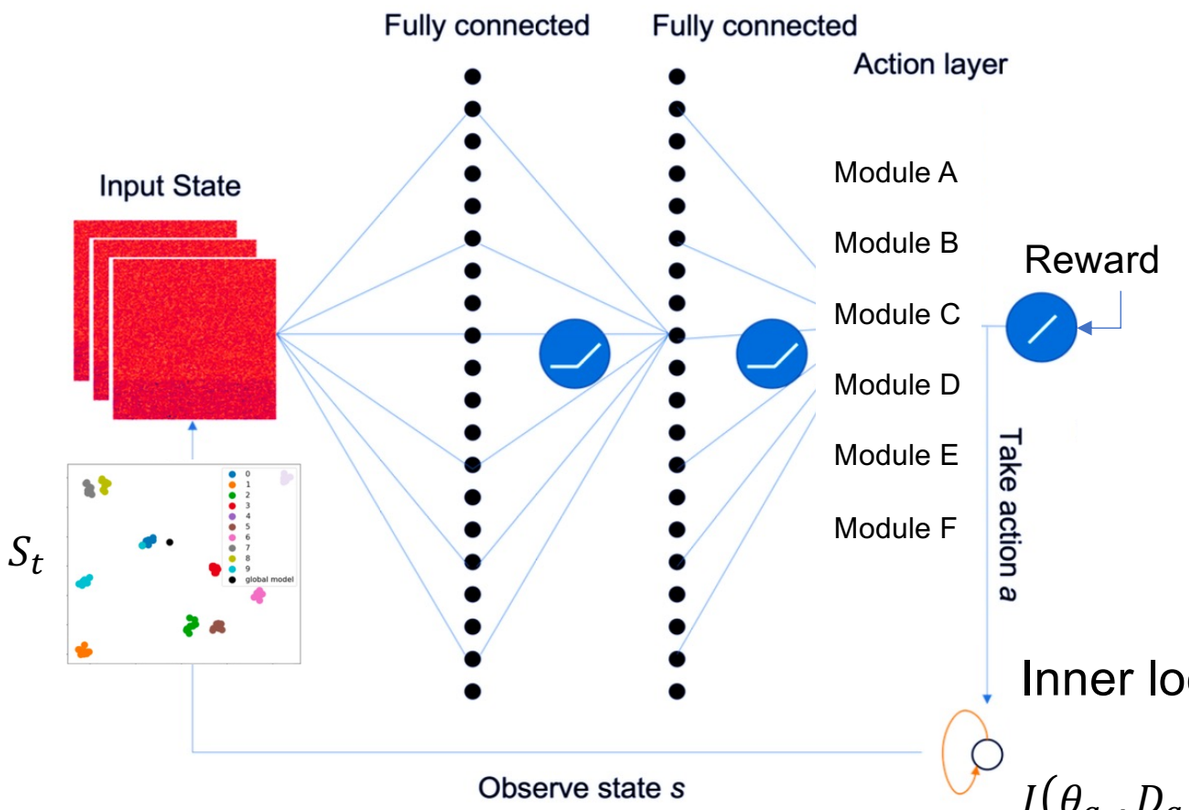
Tackling skewed class distributions as a Markov decision process



Maximize the accumulative reward

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

Meta learning policy learning with reinforcement learning



States Previous action

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

Policy Reward

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

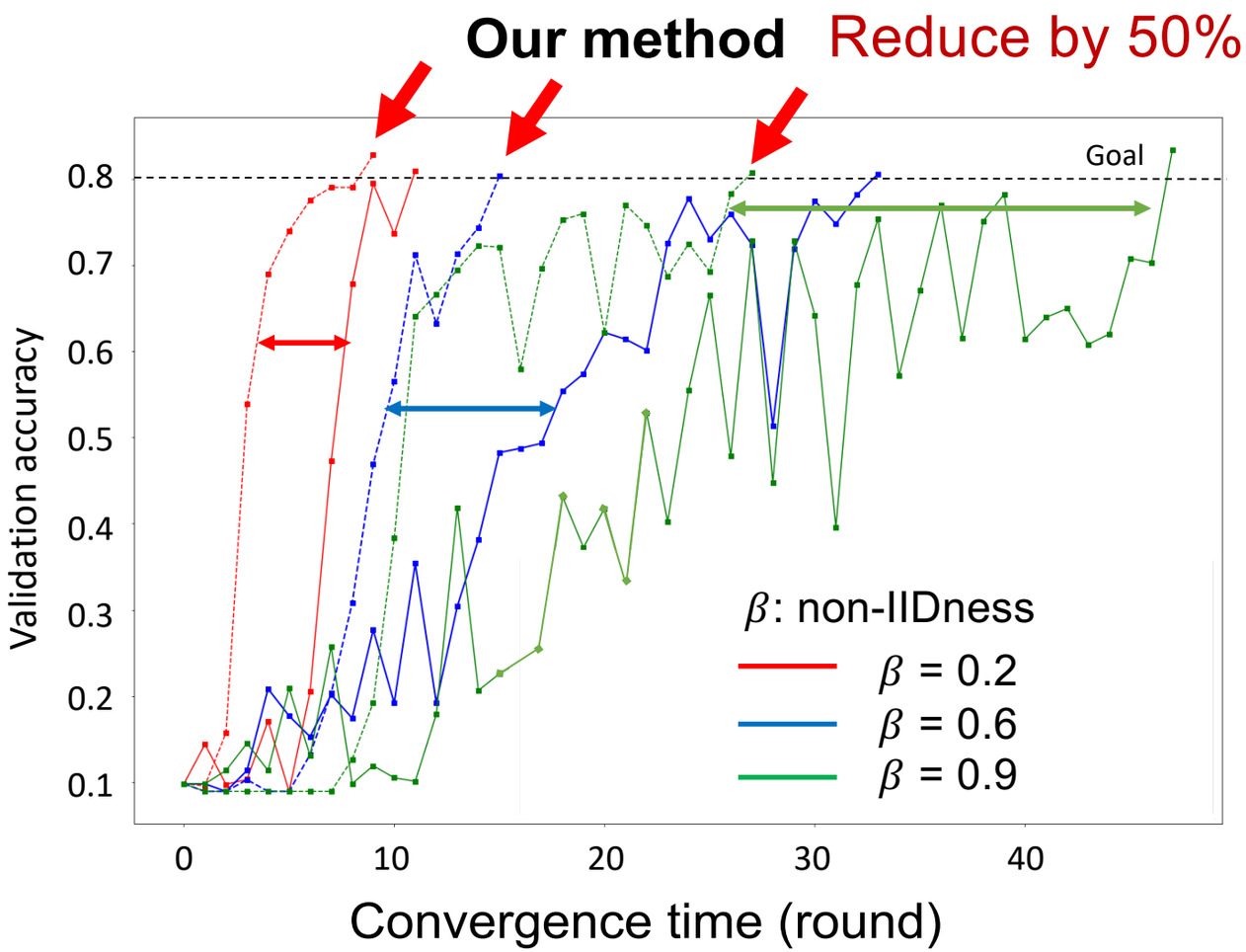
Next action

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

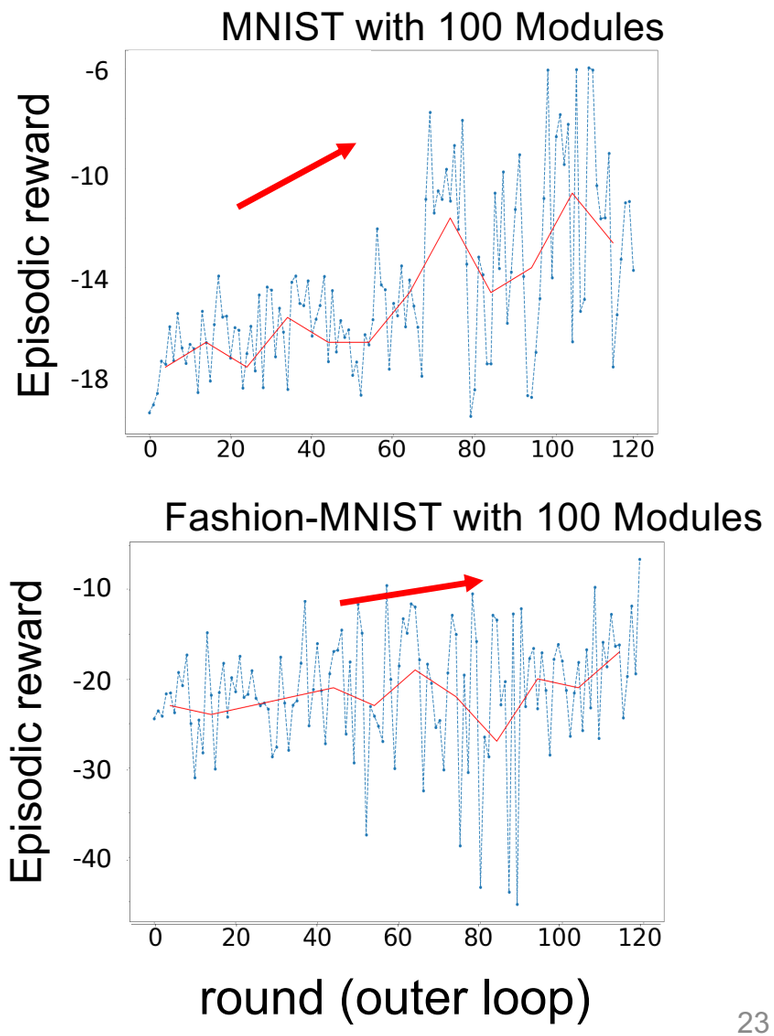
Inner loop update

$$J(\theta_{a_t}, D_{a_t}) = \frac{1}{N_{a_t}} \sum_{i=1}^{N_{a_t}} \mathcal{L}(y_i, f_{a_t}(x_i, \theta_{a_t}))$$

Convergence time reduction

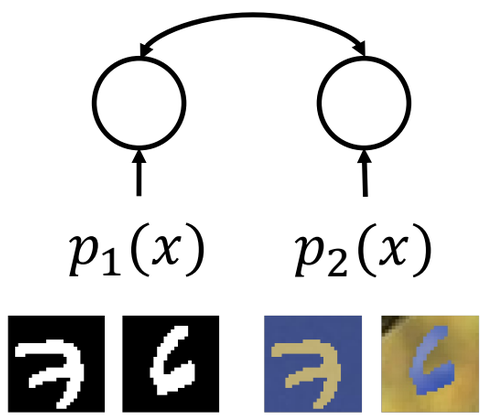


Reward curves

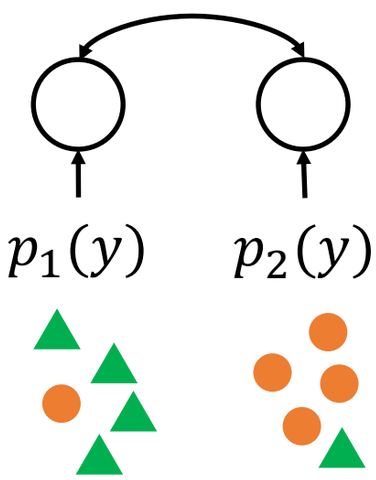


Reusable knowledge representation learning

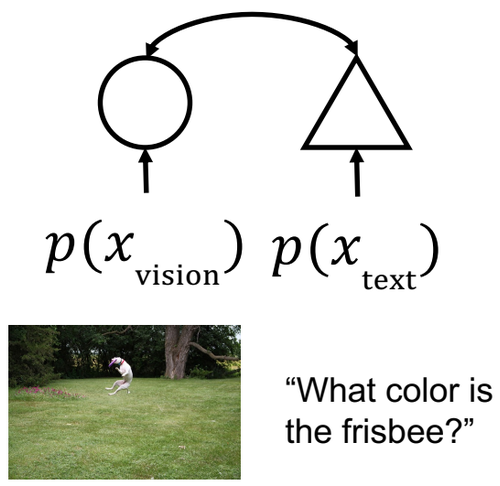
- Enhancing generalization through module knowledge transfer



(a) distributional shift



(b) class shift



(c) cross-modal

Visual Question Answering

VQA-v2 [Agrawal, 2017]

- Training: 83k images, 444k questions
- Validation: 41k images, 214k questions

Yes/No

"Are the boys sitting at a table?"

"no"

Number

"How many cats?"

"3"

Input

Output

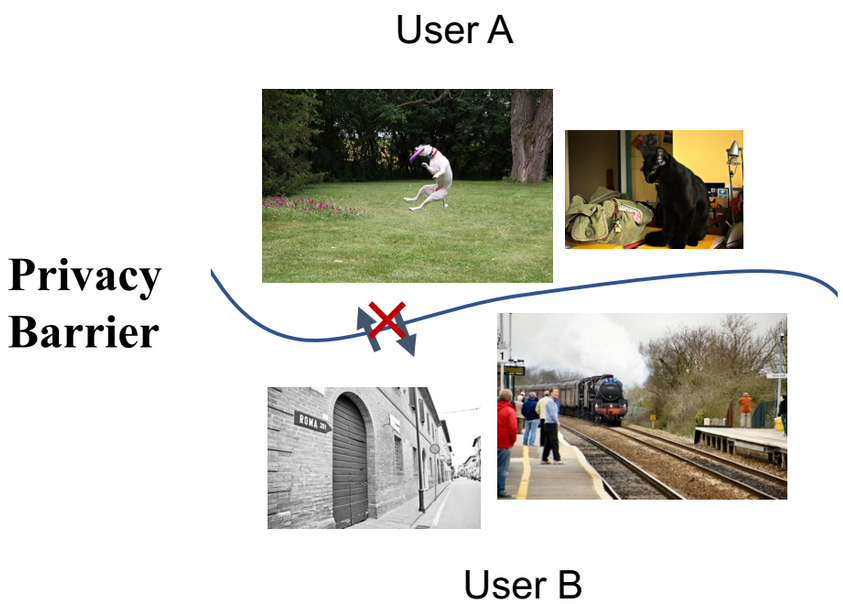
Other

"What color is the frisbee?"

"purple"

Privacy-preserving multi-client VQA

Methods	Shared Data	Shared Model	Learning Framework	Loss Function
MMNas [42]	✓	✓	Single fusion model	Cross entropy loss
QICE [46]	✓	✓	Single fusion model	Cross entropy loss + Contrastive loss
CLIP [9]	✓	✓	Single fusion model	Contrastive loss
aimNet [23]	×	✓	Federated Learning	Cross entropy loss
UniCon (Ours)	×	×	Unidirectional Split Leaning	Contrastive loss

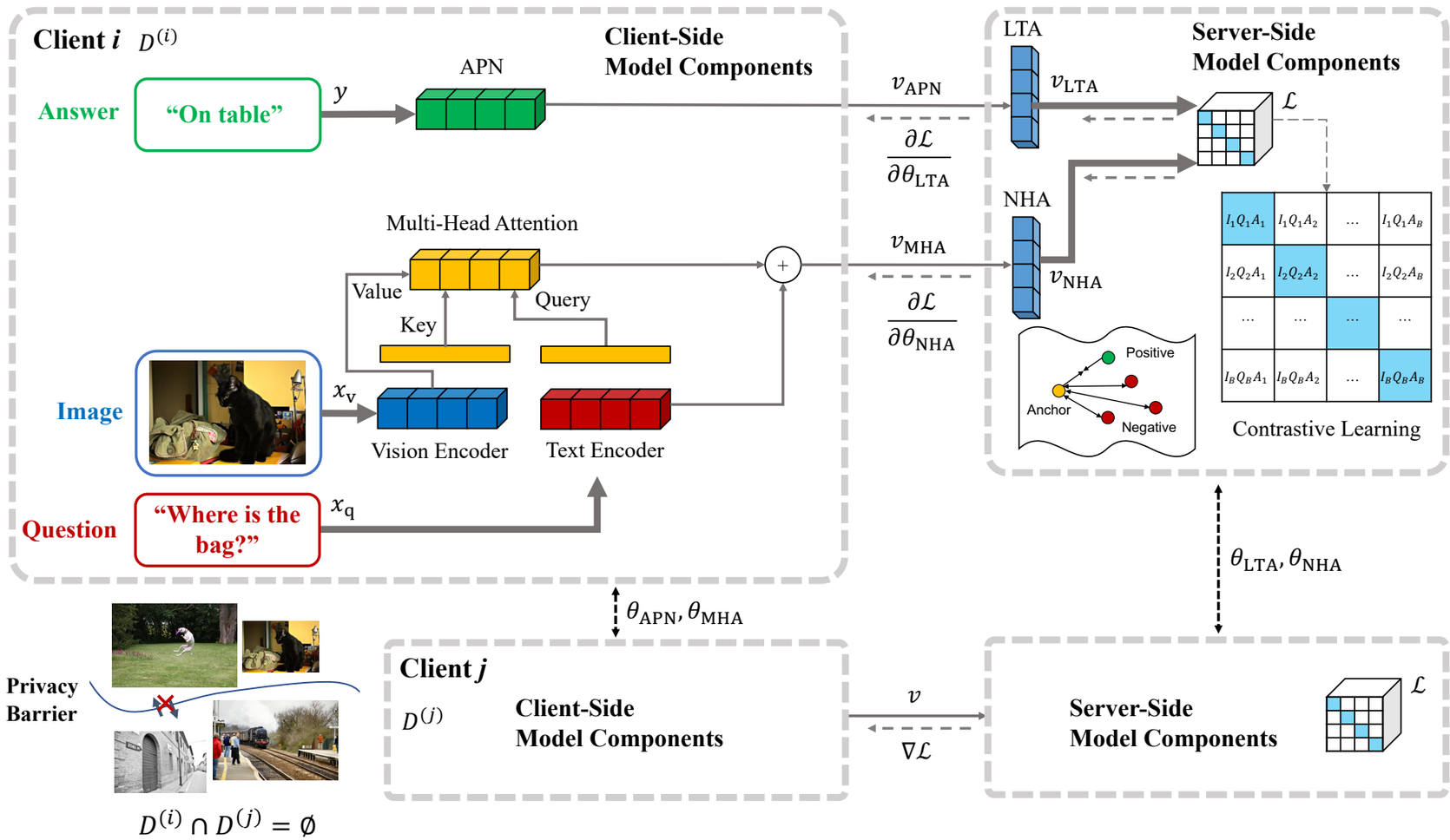


$$D^{(A)} \cap D^{(B)} = \emptyset$$

Transferring and aggregating the knowledge of these models is crucial to achieve the training goal

Sun et al. Unidirectional Split Learning with Contrastive Loss for Visual Question Answering. NeurIPS 2022 workshop.

Alignment with self-supervised learning



Learns refined representations from different clients while maintaining their privacy

Evaluation

Centralized

Transferred knowledge

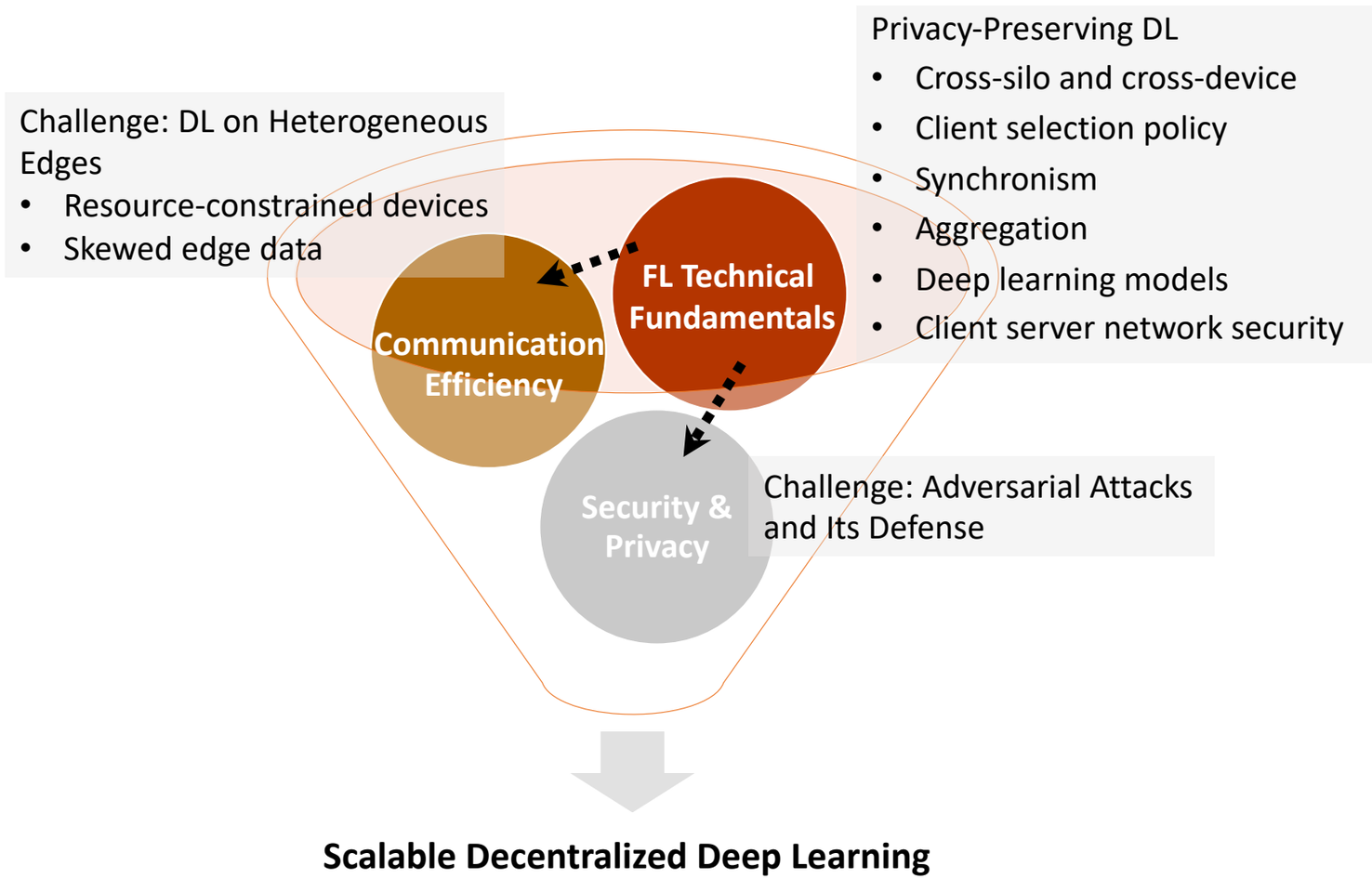
Privacy Barrier

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	**53.82**	80.06	**42.86**	**36.75**

+ Privacy guarantee

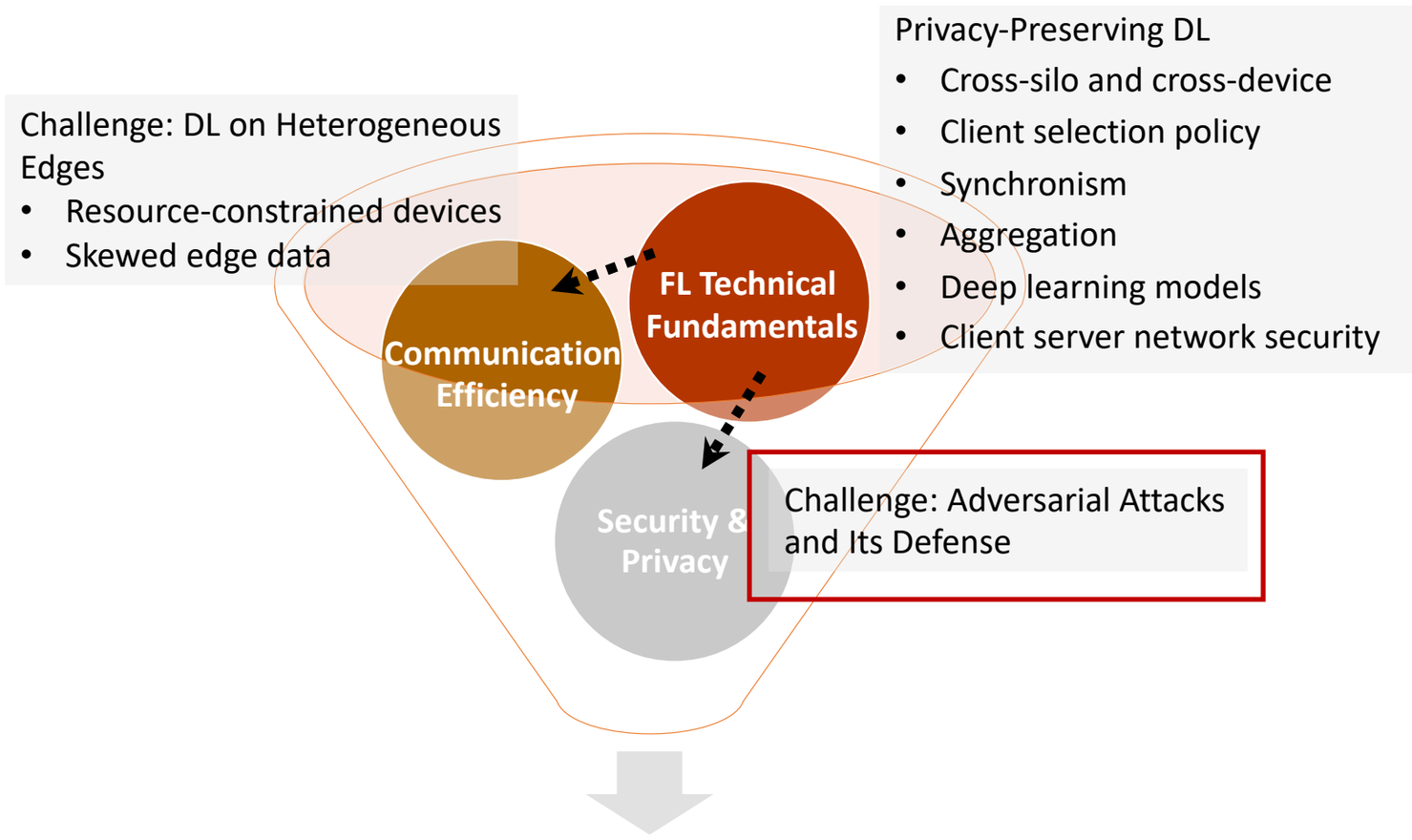
VQA Models	**UniCon** (%)			
	Overall	Yes/No	Number	Other
BAN	35.11	63.84	11.06	19.61
BUTD	40.96	66.98	13.34	28.74
MFB	42.43	68.65	23.33	27.52
MCAN-s	48.42	74.93	30.88	32.89
MCAN-l	48.44	**77.44**	30.72	32.01
MMNas-s	45.14	70.55	28.04	30.33
MMNas-l	**49.89**	74.85	**36.88**	**34.33**

Other Aspects of Scalable Decentralized ML



Sun et al. Decentralized Deep Learning for Multi-Access Edge Computing: A Survey. IEEE Transactions on Artificial Intelligence. 2022

Other Aspects of Scalable Decentralized ML



Sun et al. Decentralized Deep Learning for Multi-Access Edge Computing: A Survey. IEEE Transactions on Artificial Intelligence. 2022

Takeaways

- Knowledge sharing and coordination
- For OOD generalization and data privacy
- Interconnected models with similar architecture
- Outer-loop learning for policy optimization
- Cross-modal knowledge transfer
- A global workspace to enable collaborative ML

Meta Neural Coordination in Decentralized Neural Networks

Yuwei Sun

The University of Tokyo

RIKEN AIP

A survey paper in IEEE Transactions on AI

A summary at AAAI23

東京大学 THE UNIVERSITY OF TOKYO

AIP Center for Advanced Intelligence Project