DapperFL: Domain Adaptive Federated Learning with Model Fusion Pruning for Edge Devices

Yongzhe Jia, Xuyun Zhang, Hongsheng Hu, Kim-Kwang Raymond Choo, Lianyong Qi, Xiaolong Xu*, Amin Beheshti, Wanchun Dou

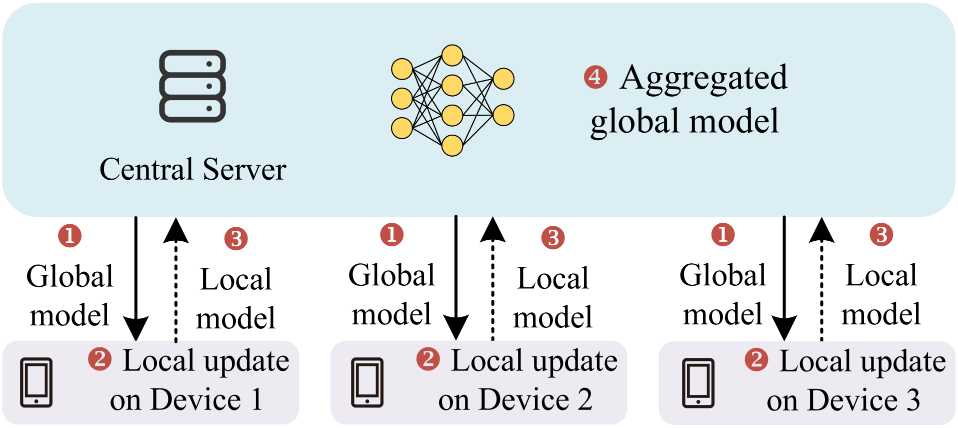
南京大學 NANJING UNIVERSITY
MACQUARIE University
THE UNIVERSITY OF NEWCASTLE AUSTRALIA
UTSA The University of Texas at San Antonio™
中国石油大学(华东) CHINA UNIVERSITY OF PETROLEUM
南京信息工程大学 Nanjing University of Information Science & Technology

NeurIPS 2024

NEURAL INFORMATION PROCESSING SYSTEMS

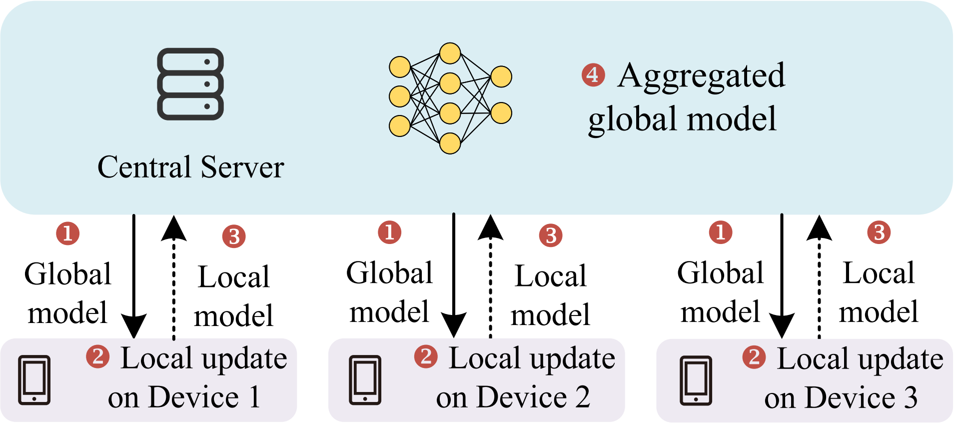
Background

Federated Learning (FL) enables participant devices (i.e., clients) to optimize their local models while a central server aggregates these local models into a global model.



Background

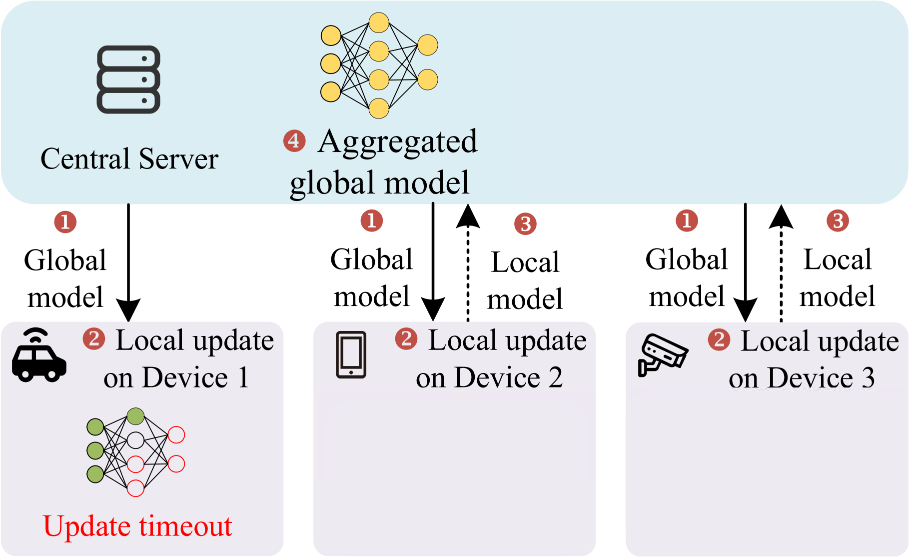
Federated Learning (FL) enables participant devices (i.e., clients) to optimize their local models while a central server aggregates these local models into a global model.



✓ Lower communication costs

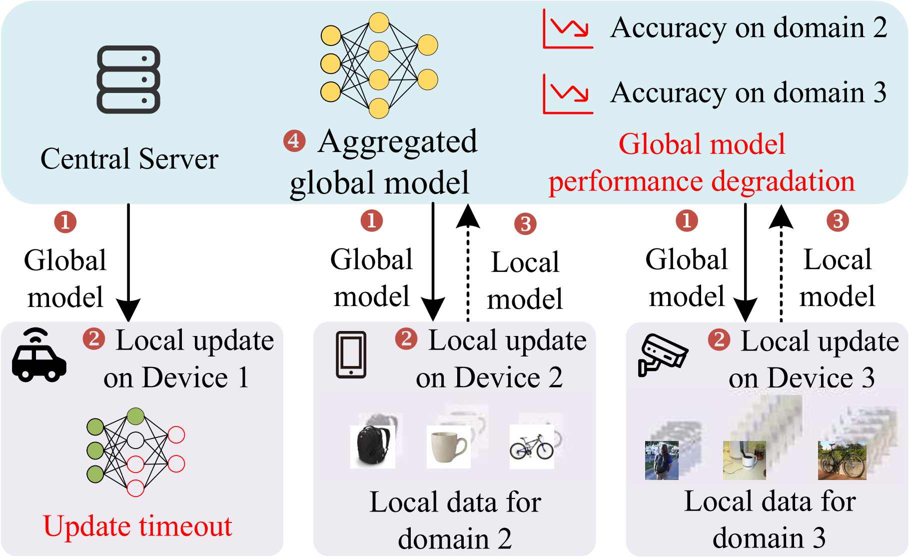
✓ Better user privacy

Motivation



✗ **System heterogeneity:**
Participant clients generally exhibit diverse and constrained system capabilities.

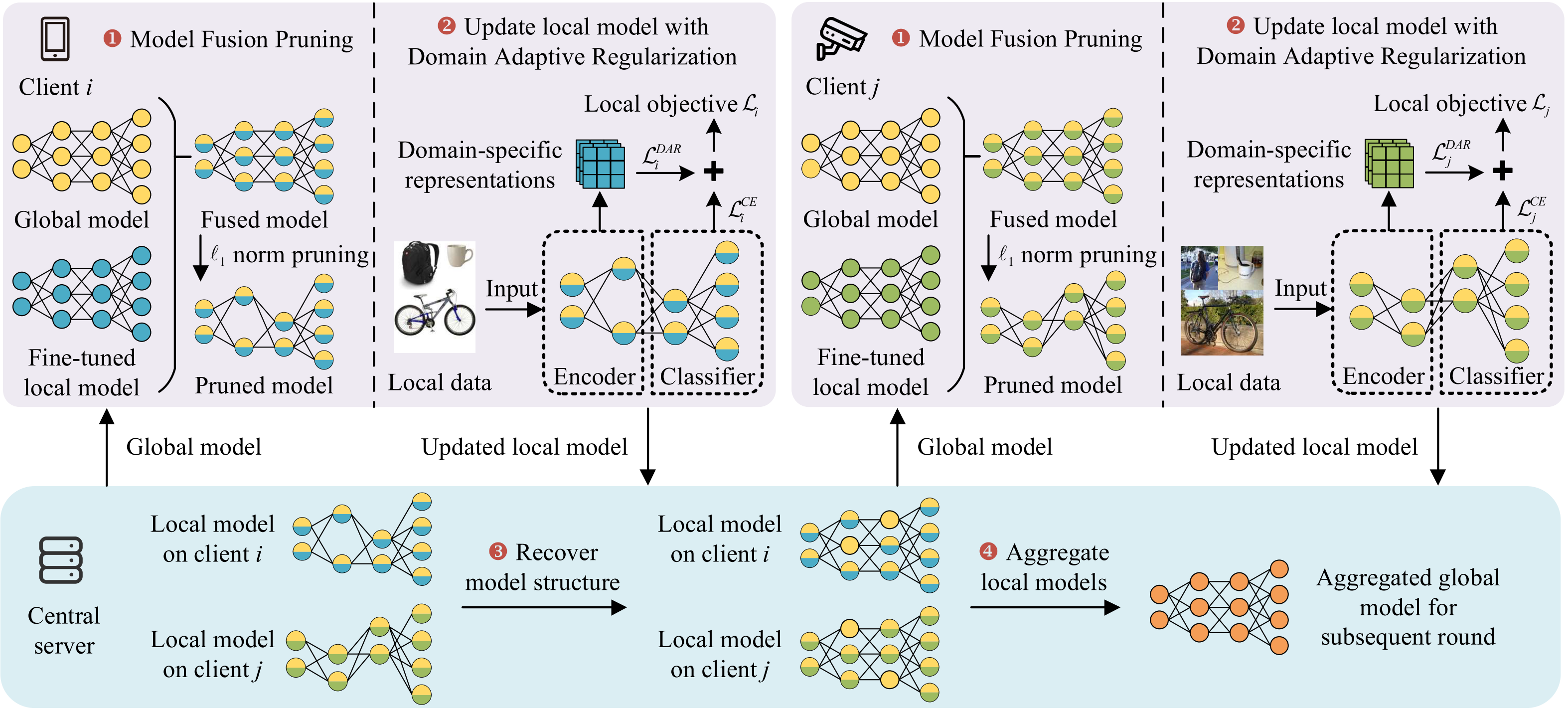
Motivation



✗ **System heterogeneity:**
Participant clients generally exhibit diverse and constrained system capabilities.

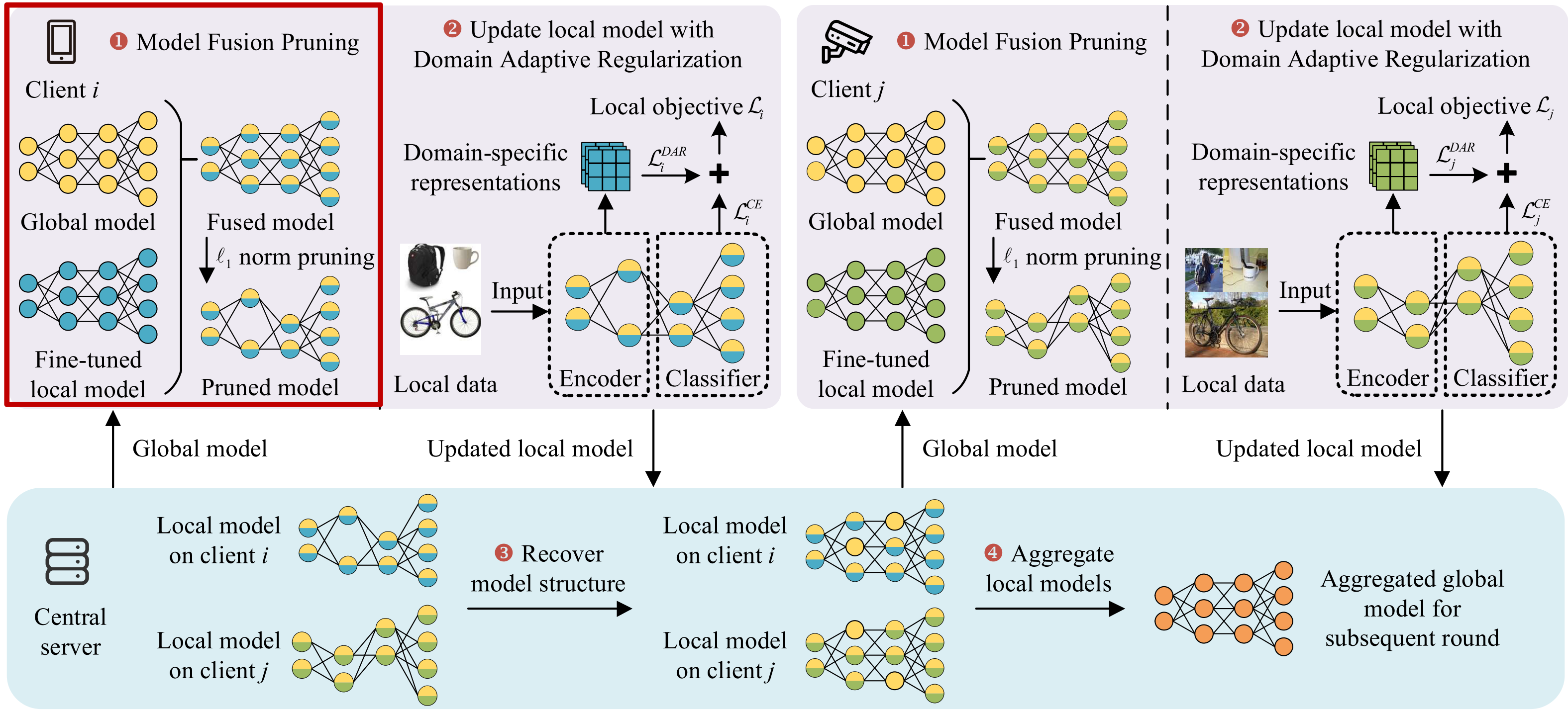
✗ **Domain shifts:**
Owing to the distributed nature of FL, the data distributions among participant clients vary significantly.

Overview



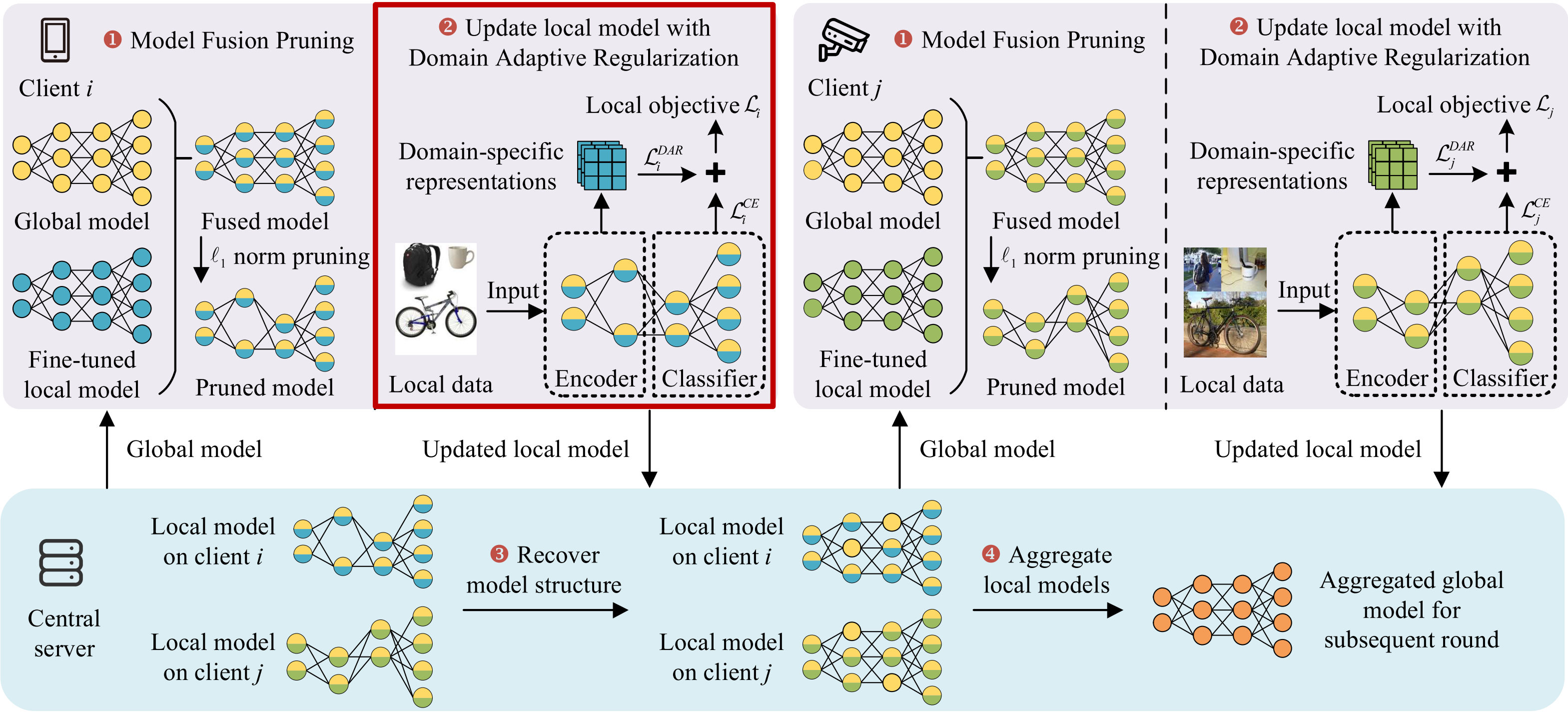
Overview of DapperFL with two clients for each communication round.

Overview



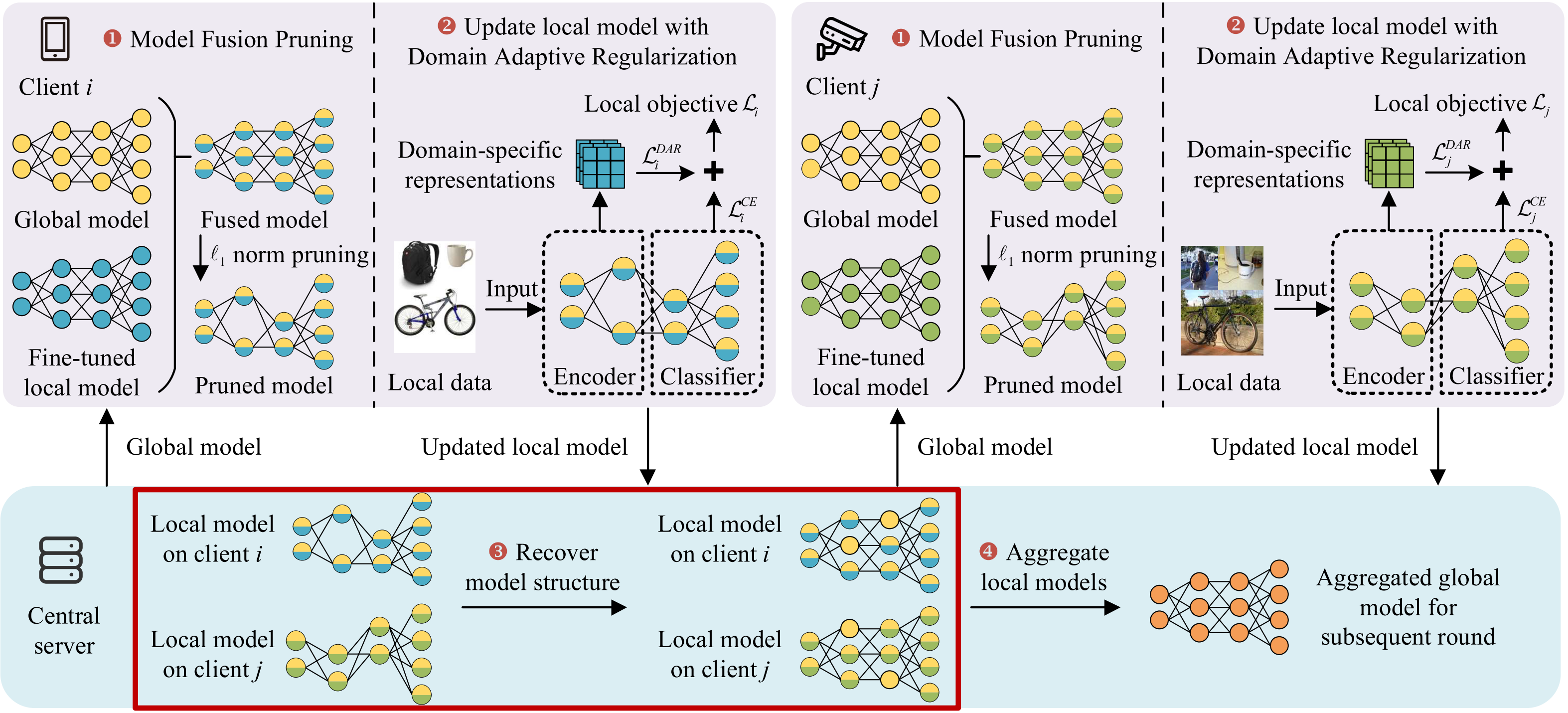
Overview of DapperFL with two clients for each communication round.

Overview



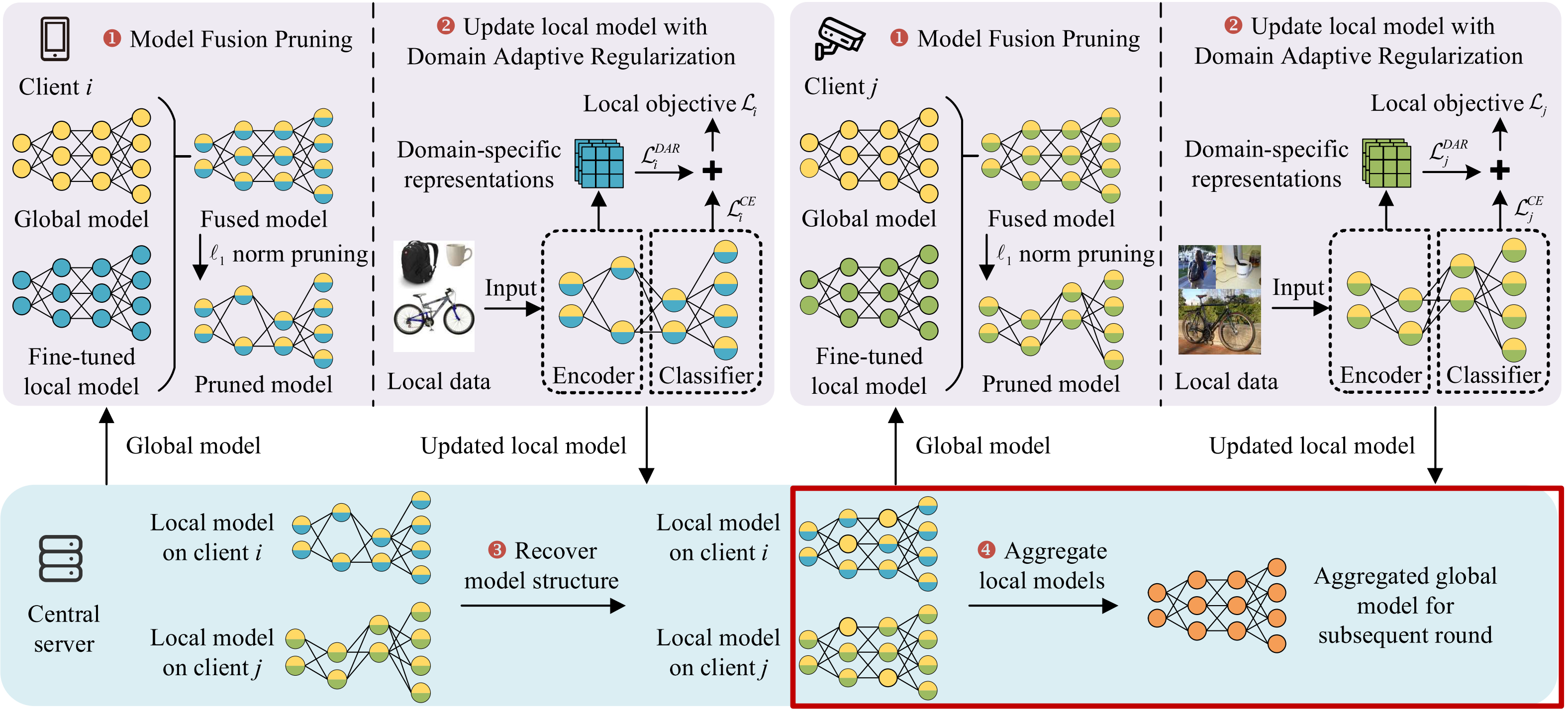
Overview of DapperFL with two clients for each communication round.

Overview



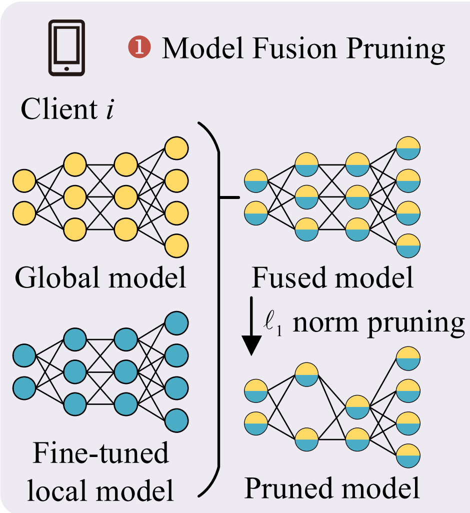
Overview of DapperFL with two clients for each communication round.

Overview



Overview of DapperFL with two clients for each communication round.

Pruning with MFP



Algorithm 1 Model Fusion Pruning of DapperFL

Input: Global model $\mathcal{W}^{t-1}$, local data $\mathcal{D}_i$, pruning ratio ρ_i

Output: Pruned local model $\boldsymbol{w}_i^t \odot \boldsymbol{M}_i^t$

$\hat{\boldsymbol{w}}_i^t \leftarrow$ Fine-tune global model $\mathcal{W}^{t-1}$ on local data $\mathcal{D}_i$

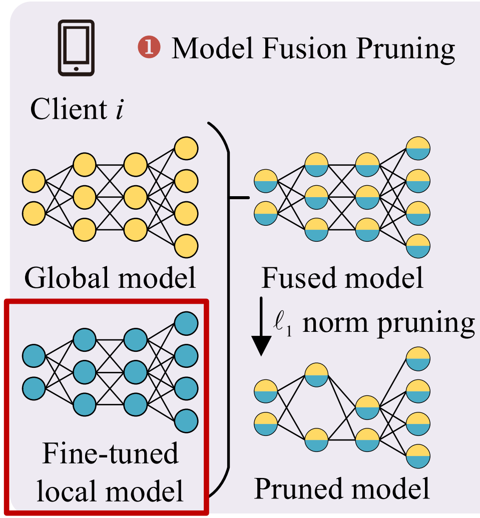
$\boldsymbol{w}_i^t \leftarrow$ Fuse the global model $\mathcal{W}^{t-1}$ into the local model $\hat{\boldsymbol{w}}_i^t$ using Eq. 1 and Eq. 2

$\boldsymbol{M}_i^t \leftarrow$ Calculate binary mask matrix by ℓ_1 norm with pruning ratio ρ_i

$\boldsymbol{w}_i^t \odot \boldsymbol{M}_i^t \leftarrow$ Prune the local model $\boldsymbol{w}_i^t$ with binary mask matrix $\boldsymbol{M}_i^t$

return Pruned local model $\boldsymbol{w}_i^t \odot \boldsymbol{M}_i^t$

Pruning with MFP



Algorithm 1 Model Fusion Pruning of DapperFL

Input: Global model $\mathcal{W}^{t-1}$, local data $\mathcal{D}_i$, pruning ratio ρ_i

Output: Pruned local model $\boldsymbol{w}_i^t \odot \boldsymbol{M}_i^t$

1: $\hat{\boldsymbol{w}}_i^t \leftarrow$ Fine-tune global model $\mathcal{W}^{t-1}$ on local data $\mathcal{D}_i$

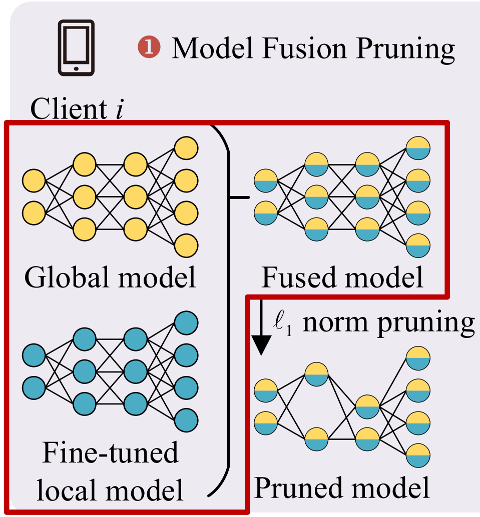
2: $\boldsymbol{w}_i^t \leftarrow$ Fuse the global model $\mathcal{W}^{t-1}$ into the local model $\hat{\boldsymbol{w}}_i^t$ using Eq. 1 and Eq. 2

3: $\boldsymbol{M}_i^t \leftarrow$ Calculate binary mask matrix by ℓ_1 norm with pruning ratio ρ_i

4: $\boldsymbol{w}_i^t \odot \boldsymbol{M}_i^t \leftarrow$ Prune the local model $\boldsymbol{w}_i^t$ with binary mask matrix $\boldsymbol{M}_i^t$

5: **return** Pruned local model $\boldsymbol{w}_i^t \odot \boldsymbol{M}_i^t$

Pruning with MFP



Algorithm 1 Model Fusion Pruning of DapperFL

Input: Global model $\mathcal{W}^{t-1}$, local data $\mathcal{D}_i$, pruning ratio ρ_i

Output: Pruned local model $\boldsymbol{w}_i^t \odot \boldsymbol{M}_i^t$

1: $\hat{\boldsymbol{w}}_i^t \leftarrow$ Fine-tune global model $\mathcal{W}^{t-1}$ on local data $\mathcal{D}_i$

2: $\boldsymbol{w}_i^t \leftarrow$ Fuse the global model $\mathcal{W}^{t-1}$ into the local model $\hat{\boldsymbol{w}}_i^t$ using Eq. 1 and Eq. 2

3: $\boldsymbol{M}_i^t \leftarrow$ Calculate binary mask matrix by ℓ_1 norm with pruning ratio ρ_i

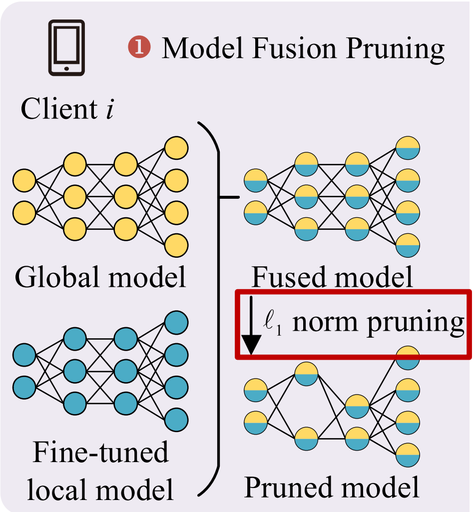
4: $\boldsymbol{w}_i^t \odot \boldsymbol{M}_i^t \leftarrow$ Prune the local model $\boldsymbol{w}_i^t$ with binary mask matrix $\boldsymbol{M}_i^t$

5: **return** Pruned local model $\boldsymbol{w}_i^t \odot \boldsymbol{M}_i^t$

Eq.1: $$\boldsymbol{w}_i^t = \alpha^t \mathcal{W}^{t-1} + (1-\alpha^t)\hat{\boldsymbol{w}}_i^t$$

Eq.2: $$\alpha^t = \max\{(1-\epsilon)^{t-1}\alpha_0, \alpha_{min}\}$$

Pruning with MFP



Algorithm 1 Model Fusion Pruning of DapperFL

Input: Global model $\mathcal{W}^{t-1}$, local data $\mathcal{D}_i$, pruning ratio ρ_i

Output: Pruned local model $\boldsymbol{w}_i^t \odot \boldsymbol{M}_i^t$

1: $\hat{\boldsymbol{w}}_i^t \leftarrow$ Fine-tune global model $\mathcal{W}^{t-1}$ on local data $\mathcal{D}_i$

2: $\boldsymbol{w}_i^t \leftarrow$ Fuse the global model $\mathcal{W}^{t-1}$ into the local model $\hat{\boldsymbol{w}}_i^t$ using Eq. 1 and Eq. 2

3: $\boldsymbol{M}_i^t \leftarrow$ Calculate binary mask matrix by ℓ_1 norm with pruning ratio ρ_i

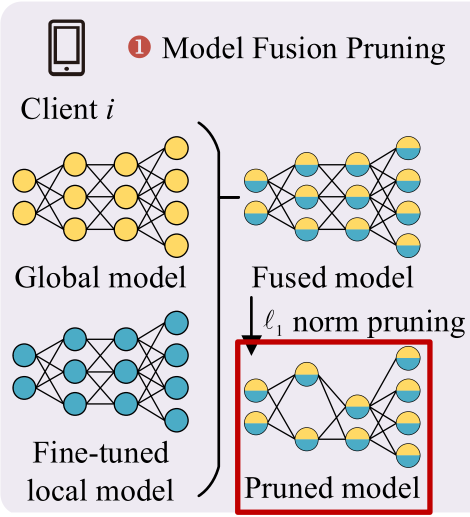
4: $\boldsymbol{w}_i^t \odot \boldsymbol{M}_i^t \leftarrow$ Prune the local model $\boldsymbol{w}_i^t$ with binary mask matrix $\boldsymbol{M}_i^t$

5: **return** Pruned local model $\boldsymbol{w}_i^t \odot \boldsymbol{M}_i^t$

Eq.1: $\boldsymbol{w}_i^t = \alpha^t \mathcal{W}^{t-1} + (1-\alpha^t)\hat{\boldsymbol{w}}_i^t$

Eq.2: $\alpha^t = \max\{(1-\epsilon)^{t-1}\alpha_0, \alpha_{min}\}$

Pruning with MFP



Algorithm 1 Model Fusion Pruning of DapperFL

Input: Global model $\mathcal{W}^{t-1}$, local data $\mathcal{D}_i$, pruning ratio ρ_i

Output: Pruned local model $\boldsymbol{w}_i^t \odot \boldsymbol{M}_i^t$

1: $\hat{\boldsymbol{w}}_i^t \leftarrow$ Fine-tune global model $\mathcal{W}^{t-1}$ on local data $\mathcal{D}_i$

2: $\boldsymbol{w}_i^t \leftarrow$ Fuse the global model $\mathcal{W}^{t-1}$ into the local model $\hat{\boldsymbol{w}}_i^t$ using Eq. 1 and Eq. 2

3: $\boldsymbol{M}_i^t \leftarrow$ Calculate binary mask matrix by ℓ_1 norm with pruning ratio ρ_i

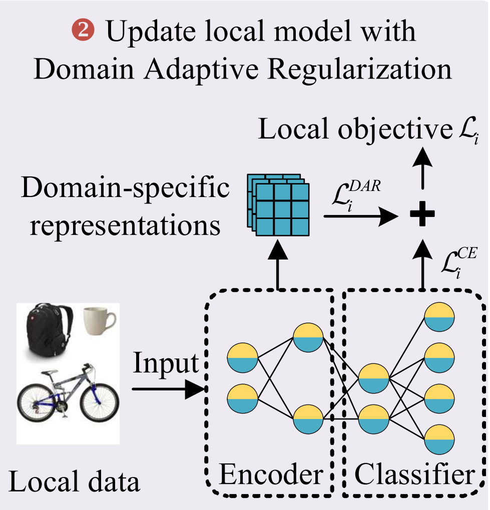
4: $\boldsymbol{w}_i^t \odot \boldsymbol{M}_i^t \leftarrow$ Prune the local model $\boldsymbol{w}_i^t$ with binary mask matrix $\boldsymbol{M}_i^t$

5: **return** Pruned local model $\boldsymbol{w}_i^t \odot \boldsymbol{M}_i^t$

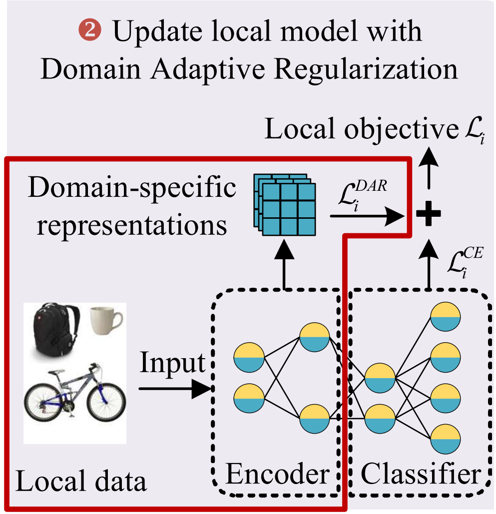
Eq.1: $\boldsymbol{w}_i^t = \alpha^t \mathcal{W}^{t-1} + (1-\alpha^t)\hat{\boldsymbol{w}}_i^t$

Eq.2: $\alpha^t = \max\{(1-\epsilon)^{t-1}\alpha_0, \alpha_{min}\}$

Updating with DAR

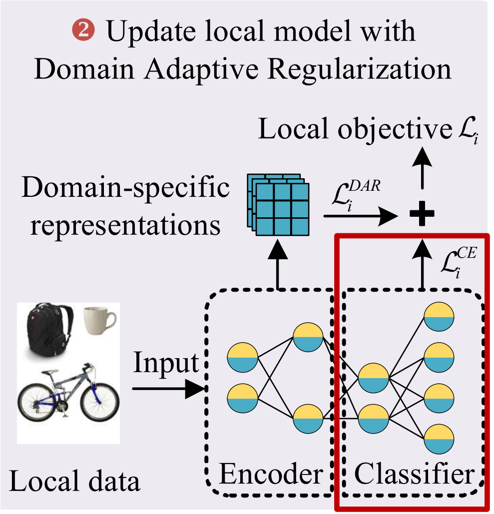


Updating with DAR



Regularization term: $\mathcal{L}_i^{DAR} = ||g_e(\boldsymbol{w}_e \odot \boldsymbol{M}_e; x_i)||_2^2$

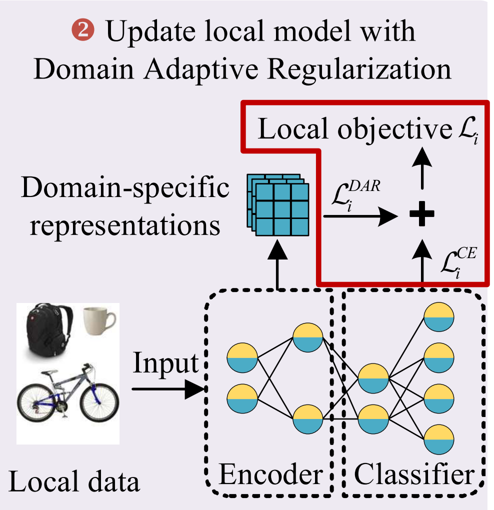
Updating with DAR



Regularization term: $\mathcal{L}_i^{DAR} = ||g_e(\boldsymbol{w}_e \odot \boldsymbol{M}_e; x_i)||_2^2$

Cross-entropy loss: $\mathcal{L}_i^{CE} = -\frac{1}{|\mathcal{K}_i|} \sum_{k \in \mathcal{K}_i} y_{i,k} \log(\hat{y}_{i,k})$

Updating with DAR

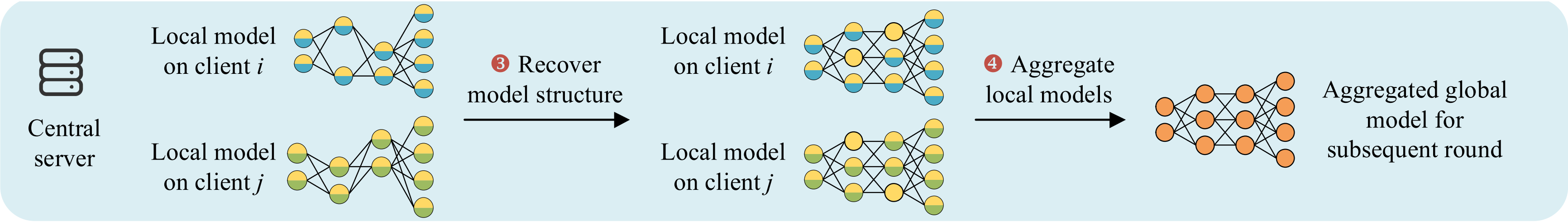


Regularization term: $\mathcal{L}_i^{DAR} = ||g_e(\boldsymbol{w}_e \odot \boldsymbol{M}_e; x_i)||_2^2$

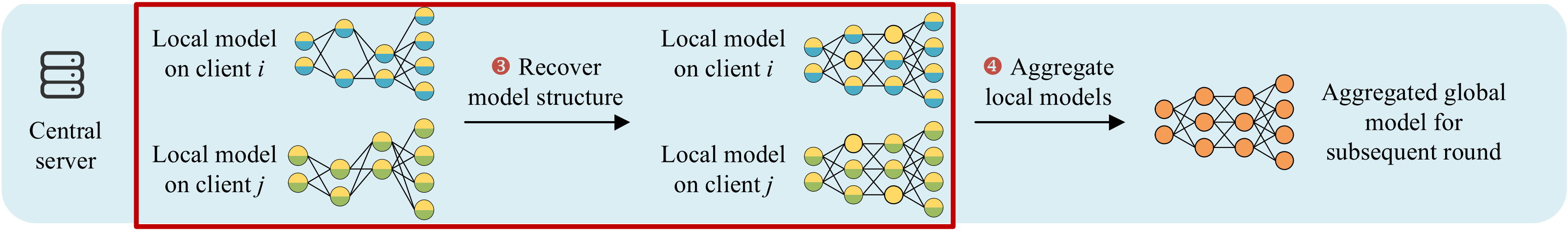
Cross-entropy loss: $\mathcal{L}_i^{CE} = -\frac{1}{|\mathcal{K}_i|} \sum_{k \in \mathcal{K}_i} y_{i,k} \log(\hat{y}_{i,k})$

Local objective: $\mathcal{L}_i = \mathcal{L}_i^{CE} + \gamma \mathcal{L}_i^{DAR}$

Model Aggregation

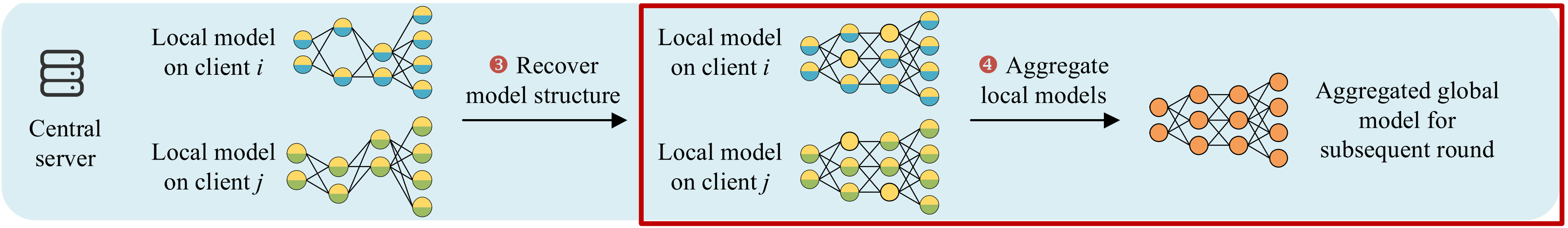


Model Aggregation



Model recovery: $$\boldsymbol{w}_i^t := \underbrace{\boldsymbol{w}_i^t \odot \boldsymbol{M}_i^t}_{\text{local knowledge}} + \underbrace{\mathcal{W}^{t-1} \odot \overline{\boldsymbol{M}}_i^t}_{\text{global knowledge}}$$

Model Aggregation



Model recovery: $$\boldsymbol{w}_i^t := \underbrace{\boldsymbol{w}_i^t \odot \boldsymbol{M}_i^t}_{\text{local knowledge}} + \underbrace{\mathcal{W}^{t-1} \odot \overline{\boldsymbol{M}}_i^t}_{\text{global knowledge}}$$

Aggregation: $$\mathcal{W}^t = \sum_{i \in \mathcal{C}} \frac{|\mathcal{D}_i|}{|\mathcal{D}|} \boldsymbol{w}_i^t$$

Accuracy across Domains

NEURAL INFORMATION PROCESSING SYSTEMS

Comparison of model accuracy on Digits:

FL frameworks	System Heter.	MNIST	USPS	SVHN	SYN	Global accuracy
FedAvg [3]	✗	95.89(1.47)	86.84(0.80)	78.39(3.24)	33.63(2.87)	71.81(0.46)
MOON [16]	✗	93.03(1.97)	78.38(5.81)	84.45(7.55)	25.97(3.28)	69.44(0.53)
FedSR [14]	✗	96.77(0.73)	86.15(2.38)	81.48(1.77)	31.64(0.40)	73.89(0.57)
FPL [15]	✗	95.54(1.78)	87.69(0.98)	83.74(4.26)	34.73(1.53)	74.17(0.95)
FedDrop [10]	✓	89.48(2.56)	82.51(1.17)	72.98(0.83)	29.35(1.97)	66.85(0.93)
FedProx [17]	✓	96.68(0.96)	83.96(0.73)	76.69(3.50)	30.95(1.42)	70.74(0.52)
FedMP [11]	✓	94.16(3.32)	85.30(2.66)	81.37(1.92)	35.12(2.00)	72.29(0.89)
NeFL [12]	✓	84.98(1.07)	88.49(4.17)	78.41(2.33)	36.02(5.72)	67.64(0.30)
DapperFL (ours)	✓	96.25(2.10)	86.30(1.24)	82.45(1.72)	37.26(2.71)	**74.30(0.26)**

Comparison of model accuracy on Office Caltech:

FL frameworks	System Heter.	Caltech	Amazon	Webcam	DSLR	Global accuracy
FedAvg [3]	✗	66.07(2.46)	76.84(3.18)	65.52(4.98)	56.67(1.98)	64.54(1.10)
MOON [16]	✗	65.62(3.74)	75.79(1.69)	72.41(2.63)	53.33(1.93)	61.86(0.79)
FedSR [14]	✗	62.95(2.25)	78.95(3.29)	75.86(3.59)	50.00(3.34)	65.47(1.13)
FPL [15]	✗	63.84(3.17)	82.63(4.11)	65.52(2.63)	60.00(3.85)	65.45(1.15)
FedDrop [10]	✓	66.07(0.89)	79.47(2.30)	56.90(3.98)	53.33(6.94)	60.58(1.42)
FedProx [17]	✓	61.61(4.09)	71.05(4.98)	68.97(4.98)	46.67(1.93)	62.08(1.11)
FedMP [11]	✓	65.62(2.49)	75.79(2.43)	56.90(3.59)	66.67(3.34)	62.34(0.93)
NeFL [12]	✓	54.91(1.57)	71.05(1.61)	77.59(4.56)	66.67(3.85)	62.26(1.34)
DapperFL (ours)	✓	64.73(1.03)	81.58(3.29)	74.14(1.99)	66.67(3.85)	**67.75(0.97)**

Accuracy across Domains

Comparison of model accuracy on Digits:

FL frameworks	System Heter.	MNIST	USPS	SVHN	SYN	Global accuracy
FedAvg [3]	✗	95.89(1.47)	86.84(0.80)	78.39(3.24)	33.63(2.87)	71.81(0.46)
MOON [16]	✗	93.03(1.97)	78.38(5.81)	84.45(7.55)	25.97(3.28)	69.44(0.53)
FedSR [14]	✗	96.77(0.73)	86.15(2.38)	81.48(1.77)	31.64(0.40)	73.89(0.57)
FPL [15]	✗	95.54(1.78)	87.69(0.98)	83.74(4.26)	34.73(1.53)	74.17(0.95)
FedDrop [10]	✓	89.48(2.56)	82.51(1.17)	72.98(0.83)	29.35(1.97)	66.85(0.93)
FedProx [17]	✓	96.68(0.96)	83.96(0.73)	76.69(3.50)	30.95(1.42)	70.74(0.52)
FedMP [11]	✓	94.16(3.32)	85.30(2.66)	81.37(1.92)	35.12(2.00)	72.29(0.89)
NeFL [12]	✓	84.98(1.07)	88.49(4.17)	78.41(2.33)	36.02(5.72)	67.64(0.30)
DapperFL (ours)	✓	96.25(2.10)	86.30(1.24)	82.45(1.72)	37.26(2.71)	**74.30(0.26)**

Comparison of model accuracy on Office Caltech:

FL frameworks	System Heter.	Caltech	Amazon	Webcam	DSLR	Global accuracy
FedAvg [3]	✗	66.07(2.46)	76.84(3.18)	65.52(4.98)	56.67(1.98)	64.54(1.10)
MOON [16]	✗	65.62(3.74)	75.79(1.69)	72.41(2.63)	53.33(1.93)	61.86(0.79)
FedSR [14]	✗	62.95(2.25)	78.95(3.29)	75.86(3.59)	50.00(3.34)	65.47(1.13)
FPL [15]	✗	63.84(3.17)	82.63(4.11)	65.52(2.63)	60.00(3.85)	65.45(1.15)
FedDrop [10]	✓	66.07(0.89)	79.47(2.30)	56.90(3.98)	53.33(6.94)	60.58(1.42)
FedProx [17]	✓	61.61(4.09)	71.05(4.98)	68.97(4.98)	46.67(1.93)	62.08(1.11)
FedMP [11]	✓	65.62(2.49)	75.79(2.43)	56.90(3.59)	66.67(3.34)	62.34(0.93)
NeFL [12]	✓	54.91(1.57)	71.05(1.61)	77.59(4.56)	66.67(3.85)	62.26(1.34)
DapperFL (ours)	✓	64.73(1.03)	81.58(3.29)	74.14(1.99)	66.67(3.85)	**67.75(0.97)**

Ablation Study

Effect of pruning ratio ρ:

Pruning ratio ρ	#Para	FLOPs	MNIST	USPS	SVHN	SYN	Global accuracy
0.2	3.92M	203.34M	94.86%	83.36%	85.55%	32.84%	73.06%
0.4	2.94M	152.50M	89.42%	80.77%	84.13%	35.57%	71.76%
0.6	1.96M	101.67M	91.79%	82.16%	77.59%	29.65%	69.27%
0.8	0.98M	50.83M	63.38%	66.17%	58.38%	21.64%	48.14%

Pruning ratio ρ	#Para	FLOPs	Caltech	Amazon	Webcam	DSLR	Global accuracy
0.2	8.94M	366.13M	68.30%	80.53%	63.79%	60.00%	66.80%
0.4	6.70M	274.60M	70.09%	79.47%	67.24%	56.67%	67.71%
0.6	4.47M	183.06M	58.48%	80.00%	67.24%	50.00%	61.02%
0.8	2.23M	91.53M	43.75%	53.16%	31.03%	33.33%	38.28%

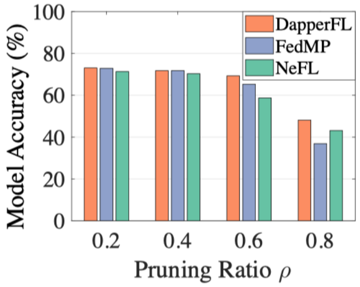
Ablation Study

Effect of pruning ratio ρ:

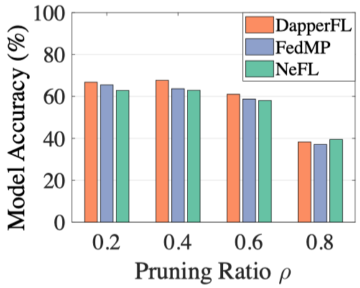
Pruning ratio ρ	#Para	FLOPs	MNIST	USPS	SVHN	SYN	Global accuracy
0.2	3.92M	203.34M	94.86%	83.36%	85.55%	32.84%	73.06%
0.4	2.94M	152.50M	89.42%	80.77%	84.13%	35.57%	71.76%
0.6	1.96M	101.67M	91.79%	82.16%	77.59%	29.65%	69.27%
0.8	0.98M	50.83M	63.38%	66.17%	58.38%	21.64%	48.14%

Pruning ratio ρ	#Para	FLOPs	Caltech	Amazon	Webcam	DSLR	Global accuracy
0.2	8.94M	366.13M	68.30%	80.53%	63.79%	60.00%	66.80%
0.4	6.70M	274.60M	70.09%	79.47%	67.24%	56.67%	67.71%
0.6	4.47M	183.06M	58.48%	80.00%	67.24%	50.00%	61.02%
0.8	2.23M	91.53M	43.75%	53.16%	31.03%	33.33%	38.28%

Comparison of model accuracy with different ρ:



(a) Digits



(b) Office Caltech

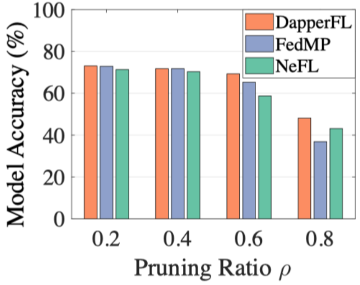
Ablation Study

Effect of pruning ratio ρ:

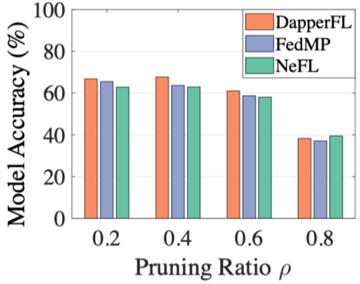
Pruning ratio ρ	#Para	FLOPs	MNIST	USPS	SVHN	SYN	Global accuracy
0.2	3.92M	203.34M	94.86%	83.36%	85.55%	32.84%	73.06%
0.4	2.94M	152.50M	89.42%	80.77%	84.13%	35.57%	71.76%
0.6	1.96M	101.67M	91.79%	82.16%	77.59%	29.65%	69.27%
0.8	0.98M	50.83M	63.38%	66.17%	58.38%	21.64%	48.14%

Pruning ratio ρ	#Para	FLOPs	Caltech	Amazon	Webcam	DSLR	Global accuracy
0.2	8.94M	366.13M	68.30%	80.53%	63.79%	60.00%	66.80%
0.4	6.70M	274.60M	70.09%	79.47%	67.24%	56.67%	67.71%
0.6	4.47M	183.06M	58.48%	80.00%	67.24%	50.00%	61.02%
0.8	2.23M	91.53M	43.75%	53.16%	31.03%	33.33%	38.28%

Comparison of model accuracy with different ρ:

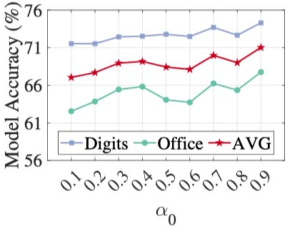


(a) Digits

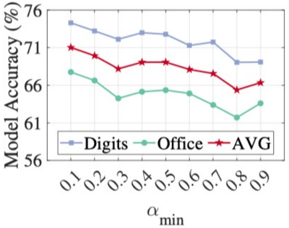


(b) Office Caltech

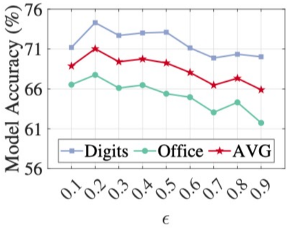
Effect of hyper-parameters in the MFP and DAR:



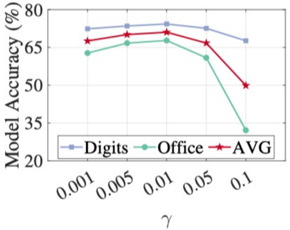
(a) Effect of α_0 in MFP



(b) Effect of α_{min} in MFP



(c) Effect of ϵ in MFP



(d) Effect of γ in DAR

Conclusion

NEURAL INFORMATION PROCESSING SYSTEMS

- We proposed the MFP module, which utilizes local and global knowledge to prune models, and we also proposed to aggregate pruned local models via a heterogeneous model aggregation algorithm.

- We proposed the DAR module, which improves the overall performance of DapperFL by implicitly encouraging pruned local models to learn robust local representations using specialized regularization techniques.

- The evaluation results show that DapperFL outperforms runner-up by up to 2.28% in terms of accuracy on two domain generalization benchmarks, while achieving adaptive model volume reduction ranging from 20% to 80%.

Thank you for your attention !