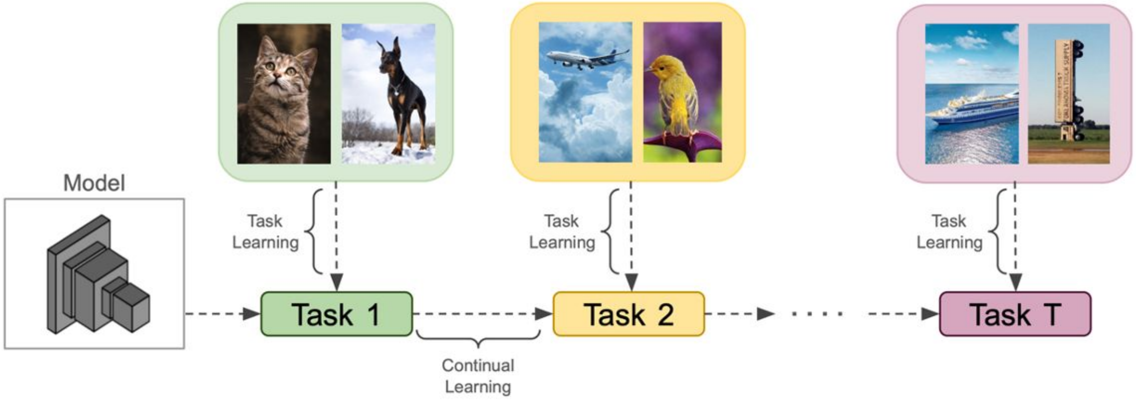
Federated Orthogonal Training: Mitigating Global Catastrophic Forgetting in Continual Federated Learning

Yavuz Bakman*, Duygu Nur Yaldiz*, Yahya Ezzeldin, Salman Avestimehr

Continual Learning



Continual Learning (CL):

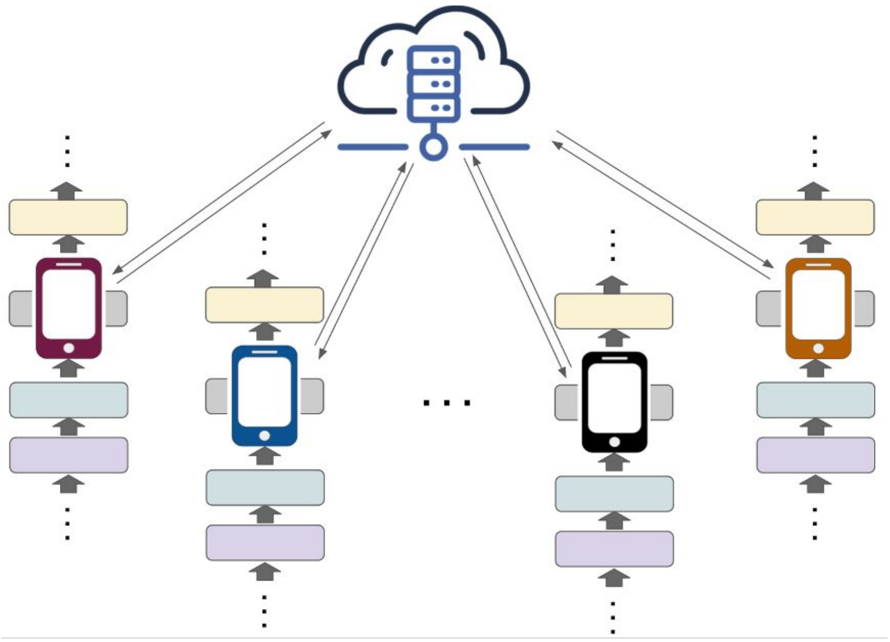
- Tasks arrive in an online manner which leads to well known phenomena *catastrophic forgetting*.
- 2 common setups: Task-incremental and Class-Incremental.

Continual Federated Learning

- In Continual Federated Learning, tasks arrive to the clients over time.

- Tasks are clearly separated from the global model perspective.

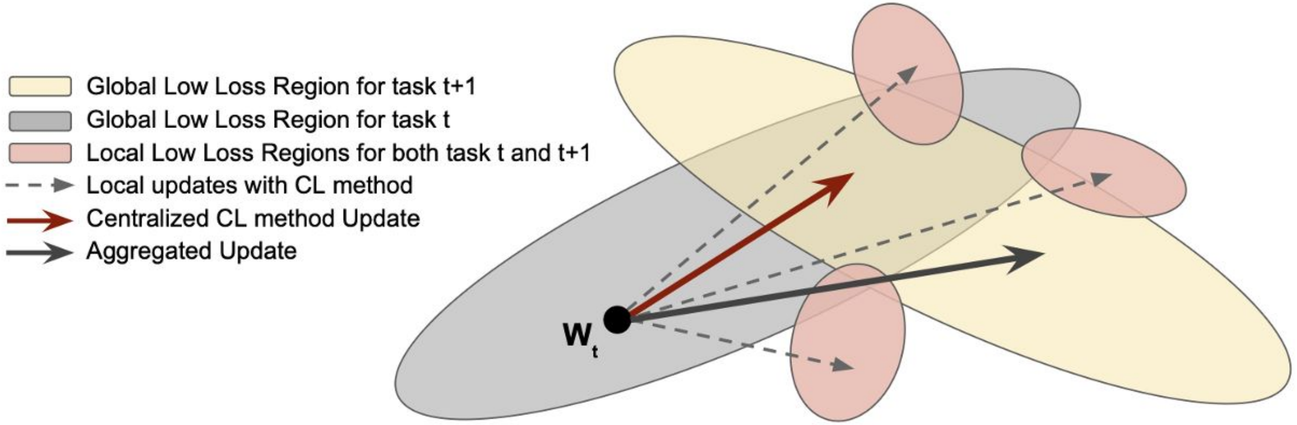
- In CFL, the global model forgets the old tasks, which is named as *global catastrophic forgetting.*



Intuition: Transferring Global Knowledge of Old Tasks

- Centralized CL algorithms prevent the disruption of knowledge acquired from previous tasks by transferring the knowledge of old tasks in current learning process.

- Naive adaptation of CL methods in CFL fail because each client transfers only its local knowledge which might not reflect the global information. This also requires storage in the client side.



Federated Orthogonal Training

Objective: Making the updates orthogonal to the old tasks' global activation subspace for each layer. Suppose the model is trained on task 1:

$$H_1^l = W_1^l X_1^l$$

where W^l is the weight of layer l , X_1^l is the input matrix of layer at task 1. Each column corresponds to one input for the layer and these inputs are distributed among the clients.

When we are training task 2, what we have is: $H_1^{l\star} = (W_1^l + \Delta W^l) X_1^l$

We want to achieve: $H_1^{l\star} = H_1^l$

Now, our goal is: $||\Delta W^l X_1^l||_F \approx 0$

Make row space of ΔW^l orthogonal to the principal column subspace of X

How can we accomplish this in Federated Learning?

Federated Orthogonal Training

Federated Projected Average: Aggregate the updates and project it to the orthogonal subspace of previous tasks.

(1) $$\delta^l_{global} \leftarrow \frac{1}{C}\sum_{i=1}^{C}\delta^l_i$$

(2) $$\delta^{l*}_{global} \leftarrow \delta^l_{global} - \mathbf{O}^l_1 {\mathbf{O}^l_1}^T \delta^l_{global}$$

(3) $$W^l \leftarrow W^l - \mu\delta^{l*}_{global}$$

$$H^l_1 = W^l_1 X^l_1$$

$$H^{l^\star}_1 = (W^l_1 + \Delta W^l)X^l_1$$

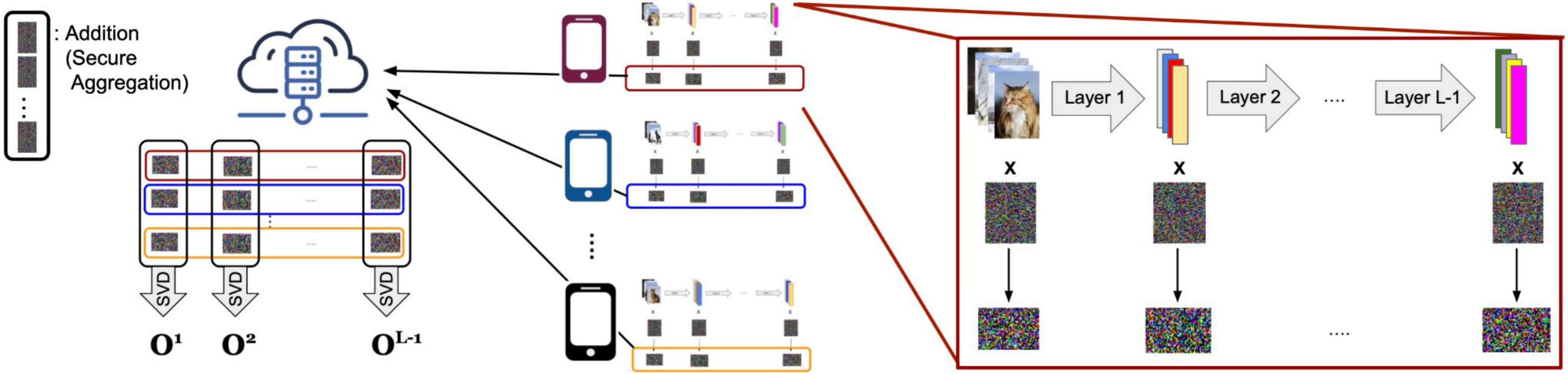
$$||\Delta W^l X^l_1||_F \approx 0$$

where δ^l_i is the local update of client i for layer l and $\mathbf{O}^l_1$ denotes global principal subspace of task 1 for layer .

How can we extract global principal subspace?

GPSE

Global Principal Subspace Extraction (GPSE)



$$\mathbf{A}_t^\ell \leftarrow \sum_{i=1}^{C} \mathbf{A}_{t,i}^\ell$$ (server side)

$$\mathbf{A}_{t,i}^\ell \leftarrow \sum_{j=1}^{n_{t,i}} \mathbf{x}_{t,i}^{j,\ell*} \mathbf{g}_j^{\ell\,T}$$ (client side)

$$\mathbf{A}_t^\ell = \sum_{i=1}^{C} \sum_{j=1}^{n_{t,i}} \mathbf{x}_{t,i}^{j,\ell*} \mathbf{g}_{j,i}^{\ell\ T} = \boxed{\mathbf{X}_t^{\ell*} \times \mathbf{G}^\ell}$$

GPSE Discussion

Communication Cost

- does not scale with the number of samples
- bounded by the model size

	Model Size	$\{\mathbf{A}^\ell\}_{\ell=1}^L$	$\{\mathbf{O}^\ell\}_{\ell=1}^L$
Resnet-18	2.56 MB	0.64 MB	0.08 MB
Alexnet	1.42 MB	0.24 MB	0.048 MB

Computation Cost

- there is no backward pass
- only forward pass and matrix multiplication

	1 Local Epoch(s)	GPSE (s)
PMNIST	0.067	0.024
CIFAR100	0.074	0.172
Mini-IMGNT	0.194	0.215

Data Heterogeneity: Obtaining principal subspace is independent of data distribution.

Privacy : GPSE results in a distributed application of the Johnson-Lindenstrauss (JL) transform, which has been shown to provide differential privacy guarantee.

Results

	Method	PMNIST		CIFAR100		5-Datasets		Mini-Imagenet	
		ACC(%)	FGT(%)	ACC(%)	FGT(%)	ACC(%)	FGT(%)	ACC(%)	FGT(%)
IID	FL	85.68±0.57	8.29±0.62	63.12±0.48	13.57±1.57	77.95±0.58	13.46±1.26	50.43±1.97	33.08±1.96
	EWC+FL	86.74±0.97	8.84±1.40	63.13±0.65	13.48±1.77	77.68±0.55	13.76±1.74	47.00±1.46	36.68±1.57
	ER+FL	88.61±0.50	6.62±0.06	65.42±0.49	11.72±0.39	80.01±1.17	10.11±2.68	55.26±2.95	27.80±3.23
	RGO+FL	89.81±0.20	4.84±0.55	63.91±0.53	14.39±0.54	84.86±1.24	3.47±0.52	51.00±1.88	31.76±1.87
	GPM+FL	88.27±0.75	6.88±0.63	59.43±0.48	18.13±0.59	80.28±1.02	10.7±0.61	50.26±1.90	30.04±1.22
	GLFC	83.76±1.05	14.70±1.13	65.16±0.48	11.74±0.59	81.64±0.22	10.07±0.61	59.26±1.23	23.73±1.27
	FOT(Ours)	**90.35**±0.06	**1.75**±0.06	**71.90**±0.06	**0.87**±0.10	**85.21**±0.93	**1.11**±0.31	**69.07**±0.73	**0.19**±0.11
non-IID	FL	80.06±0.43	9.21±0.67	62.56±0.17	10.49±0.40	67.71±7.75	21.55±9.73	41.00±0.41	32.92±1.25
	EWC+FL	81.14±1.15	10.78±0.96	62.57±0.47	10.41±0.83	68.41±3.14	21.42±4.19	41.09±2.33	32.23±1.20
	ER+FL	82.77±1.56	9.32±2.57	58.57±1.59	14.78±2.18	70.68±2.46	18.48±3.27	49.23±2.09	24.40±0.50
	RGO+FL	79.31±3.48	15.72±3.96	40.83±2.86	31.57±1.61	77.92±0.80	6.89±1.12	40.43±0.23	33.24±0.80
	GPM+FL	72.17±1.22	22.6±2.54	34.15±2.31	34.60±2.87	72.21±0.97	16.84±2.15	29.63±3.12	40.73±4.20
	GLFC	84.06±0.24	12.13±0.54	59.89±1.15	15.36±1.69	77.80±0.49	11.06±1.15	50.41±1.85	23.46±2.21
	FOT(Ours)	**85.21**±0.13	**1.97**±0.22	**66.31**±0.25	**0.60**±0.43	**79.23**±1.02	**0.65**±0.27	**62.06**±0.59	**0.17**±0.27

$$\mathrm{ACC} = \frac{1}{K}\sum_{i=1}^{K} a_{i,K} \qquad \mathrm{FGT} = \frac{1}{K-1}\sum_{i=1}^{K-1} a_{i,i} - a_{i,K}$$