南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Model-Contrastive Federated Learning

Qinbin Li
National University of Singapore
qinbin@comp.nus.edu.sg

Bingsheng He
National University of Singapore
hebs@comp.nus.edu.sg

Dawn Song
UC Berkeley
dawnsong@berkeley.edu

Federated learning：

Federated learning enables multiple parties to collaboratively train a machine learning model without communicating their local data. A key challenge in federated learning is to handle the heterogeneity of local data distribution across parties.

Generic Federated Learning (G-FL) aims to create a single global model that performs well across all clients.

Personalized Federated Learning (P-FL) acknowledges the heterogeneity among clients and focuses on tailoring models to individual clients.

Data heterogeneity:

A key challenge in federated learning is the heterogeneity of data distribution on different parties. The data can be non-identically distributed among the parties in many real-world applications, which can degrade the performance of federated learning. Non-IID data includes forms as follow:

Feature distribution skew

Label distribution skew

Same label but different features

Same features but different label

Distribution skew

When each party updates its local model, its local objective may be far from the global objective. Thus, the averaged global model is away from the global optima.

Contrastive Learning:

The key idea of contrastive learning is to reduce the distance be tween the representations of different augmented views of the same image (i.e., positive pairs), and increase the distance between the representations of augmented views of different images (i.e., negative pairs).

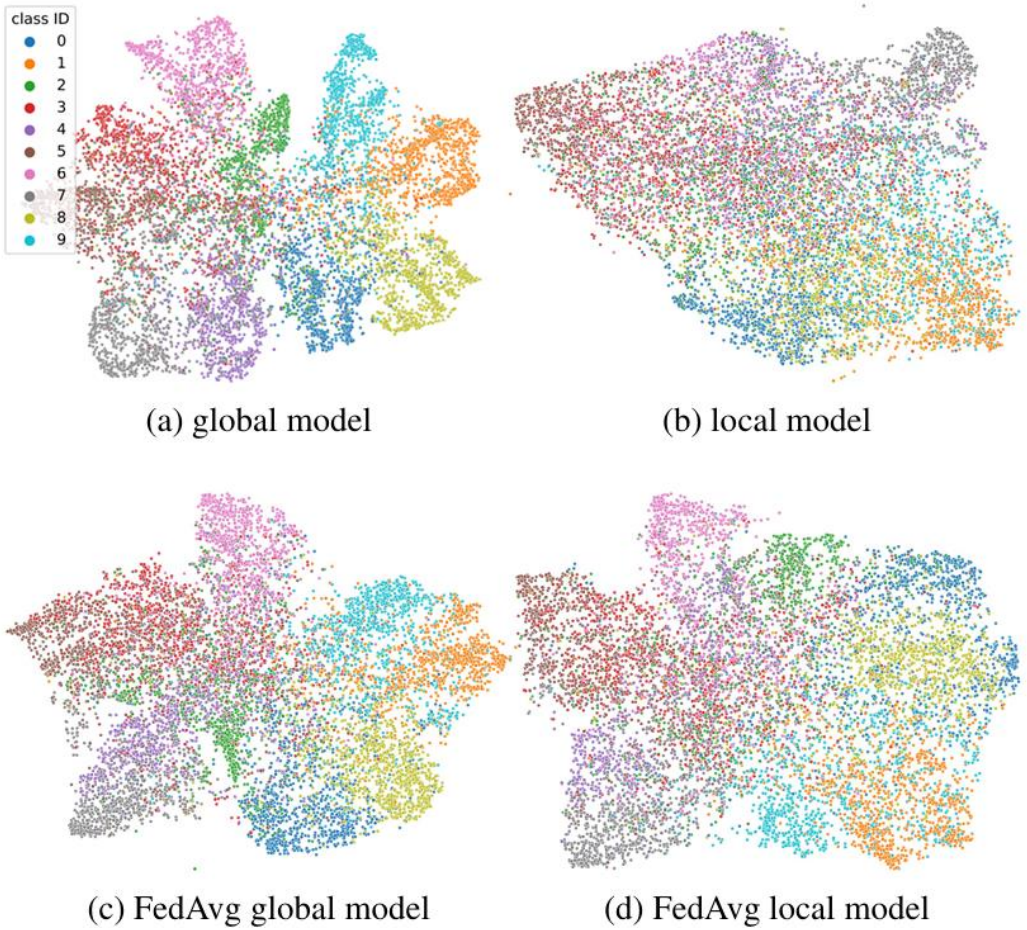
A typical contrastive learning framework is SimCLR. Given an image x, SimCLR first creates two correlated views of this image using different data augmentation operators. A base encoder and a projection head are trained to extract the representation vectors and map the representations to a latent space.

$$l_{i,j} = -\log \frac{\exp(\text{sim}(x_i, x_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{I}_{[k \neq i]} \exp(\text{sim}(x_i, x_k)/\tau)}$$

NT-Xent Loss

Model-Contrastive Federated Learning:

MOON is based on an intuitive idea: the model trained on the whole dataset is able to extract a better feature representation than the model trained on a skewed subset. Visualized the hidden vectors :



(a) global model (b) local model

(c) FedAvg global model (d) FedAvg local model

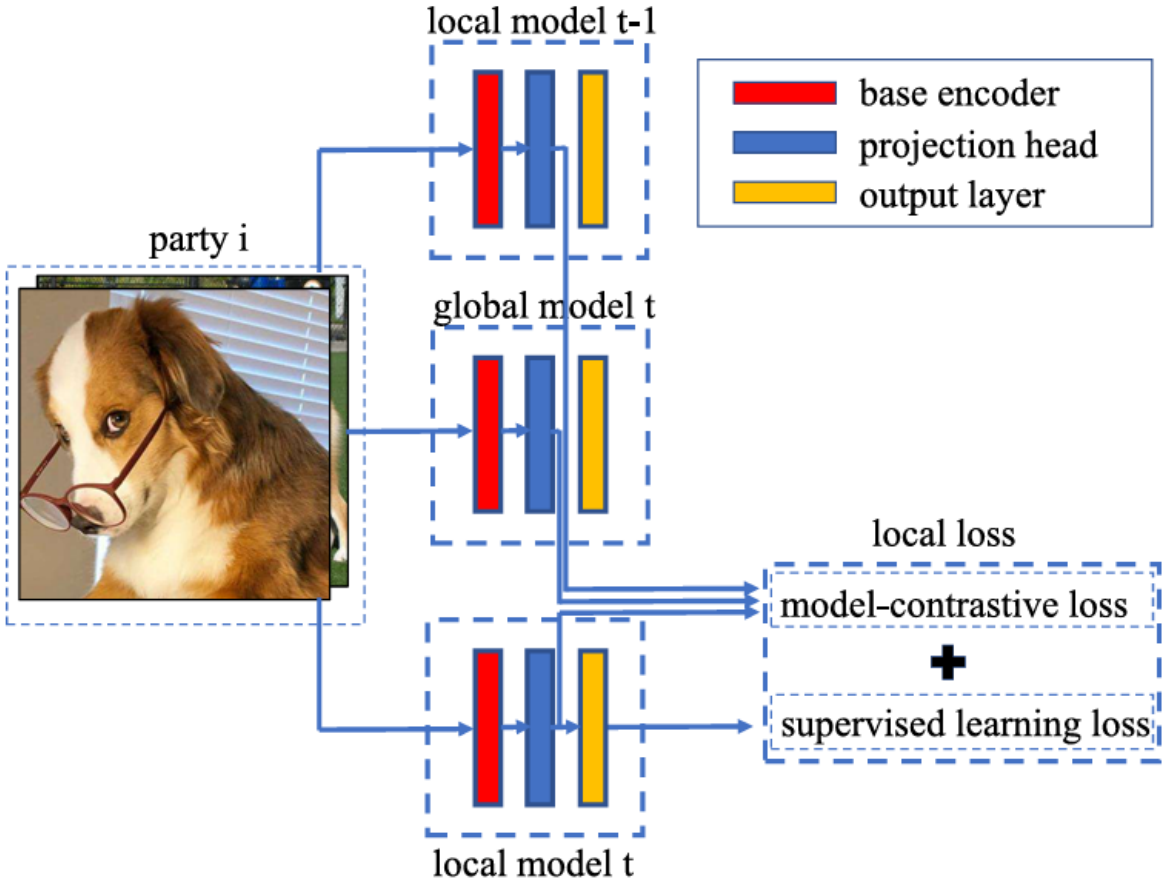
MOON : Model-Contrastive Federated Learning

MOON aims to decrease the distance between the representation learned by the local model and the representation learned by the global model, and increase the distance between the representation learned by the local model and the representation learned by the previous local model.

Network Architecture

The network has three components: a base encoder, a projection head, and an output layer.

We use Fw(·) to denote the whole network and Rw(·) to denote the network before the output layer.



Local Objective

Local loss consists two parts. The first part is a typical loss term (e.g., cross-entropy loss) in supervised learning denoted as ℓ_{sup} . The second part is our proposed model-contrastive loss term denoted as ℓ_{con} .

For every input x, we extract the representation of x from the global model w^t

$$z_{glob} = R_{w^t}(x) \qquad z_{prev} = R_{w_i^{t-1}}(x)$$

Similar to NT-Xent loss, we define model-contrastive loss as:

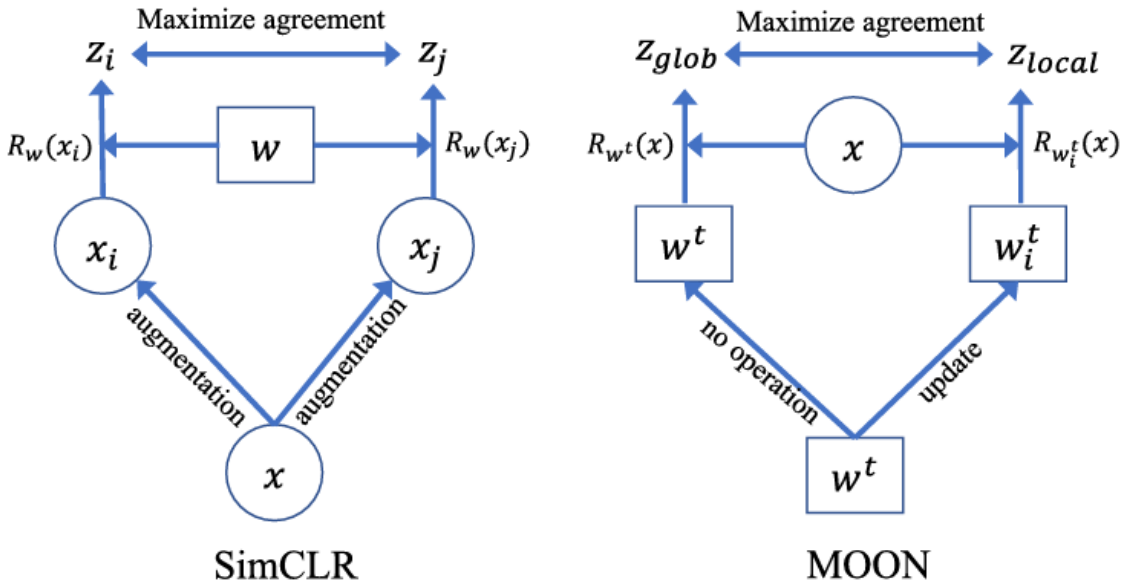
$$\ell_{con} = -\log \frac{\exp(\text{sim}(z, z_{glob})/\tau)}{\exp(\text{sim}(z, z_{glob})/\tau) + \exp(\text{sim}(z, z_{prev})/\tau)}$$

τ denotes a temperature parameter

Method

The loss of an input (x, y) is computed by

$$\ell = \ell_{sup}(w_i^t; (x, y)) + \mu \ell_{con}(w_i^t; w_i^{t-1}; w^t; x)$$



An notable thing is that considering an ideal case where the local model is good enough and learns (almost) the same representation as the global model (i.e., Zglob = Zprev), the model-contrastive loss will be a constant (i.e., −log 1/2). Thus, MOON will produce the same result as FedAvg, since there is no heterogeneity issue. In this sense, our approach is robust regardless of different amount of drifts.

Method

Algorithm 1: The MOON framework

Input: number of communication rounds T, number of parties N, number of local epochs E, temperature τ, learning rate η, hyper-parameter μ

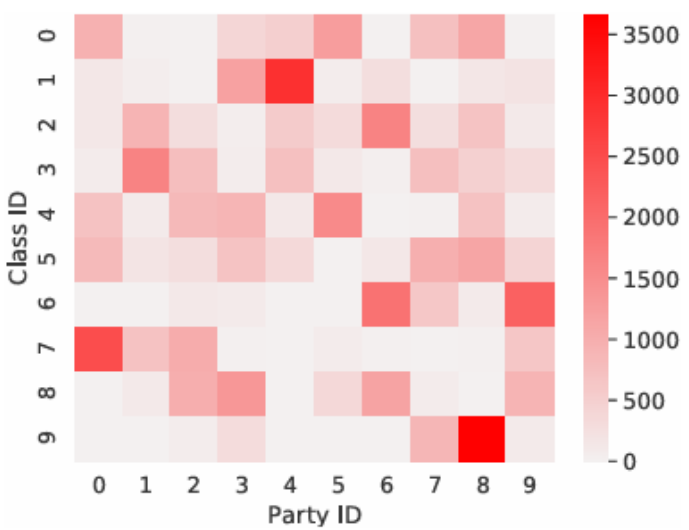
Output: The final model w^T

Server executes:
initialize w^0
for $t = 0, 1, ..., T-1$ **do**
  for $i = 1, 2, ..., N$ **in parallel do**
    send the global model w^t to P_i
    $w_i^t \leftarrow$ **PartyLocalTraining**(i, w^t)
  $w^{t+1} \leftarrow \sum_{k=1}^{N} \frac{|\mathcal{D}^i|}{|\mathcal{D}|} w_k^t$
return w^T

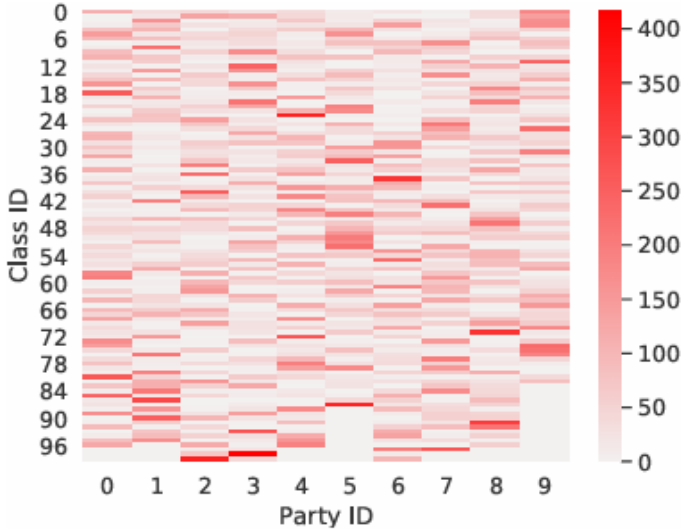
PartyLocalTraining(i, w^t):
$w_i^t \leftarrow w^t$
for epoch $i = 1, 2, ..., E$ **do**
  for each batch $\mathbf{b} = \{x, y\}$ of $\mathcal{D}^i$ **do**
    $\ell_{sup} \leftarrow CrossEntropyLoss(F_{w_i^t}(x), y)$
    $z \leftarrow R_{w_i^t}(x)$
    $z_{glob} \leftarrow R_{w^t}(x)$
    $z_{prev} \leftarrow R_{w_i^{t-1}}(x)$
    $\ell_{con} \leftarrow -\log \frac{\exp(\text{sim}(z, z_{glob})/\tau)}{\exp(\text{sim}(z, z_{glob})/\tau) + \exp(\text{sim}(z, z_{prev})/\tau)}$
    $\ell \leftarrow \ell_{sup} + \mu \ell_{con}$
    $w_i^t \leftarrow w_i^t - \eta \nabla \ell$
return w_i^t to server

Experiment

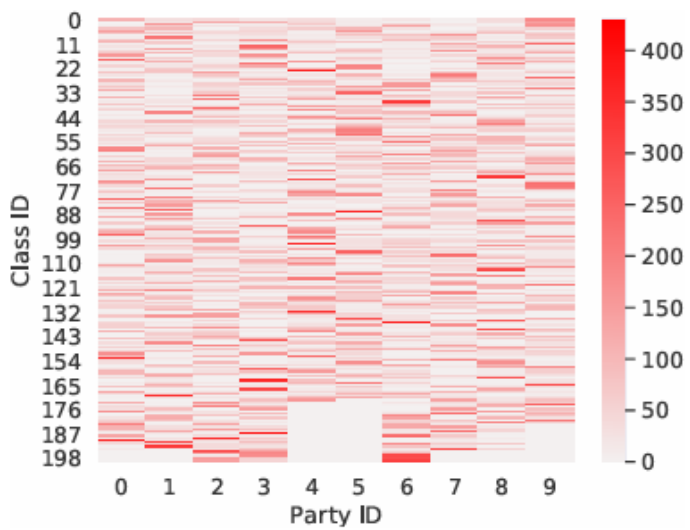
Accuracy Comparison



(a) CIFAR-10



(b) CIFAR-100

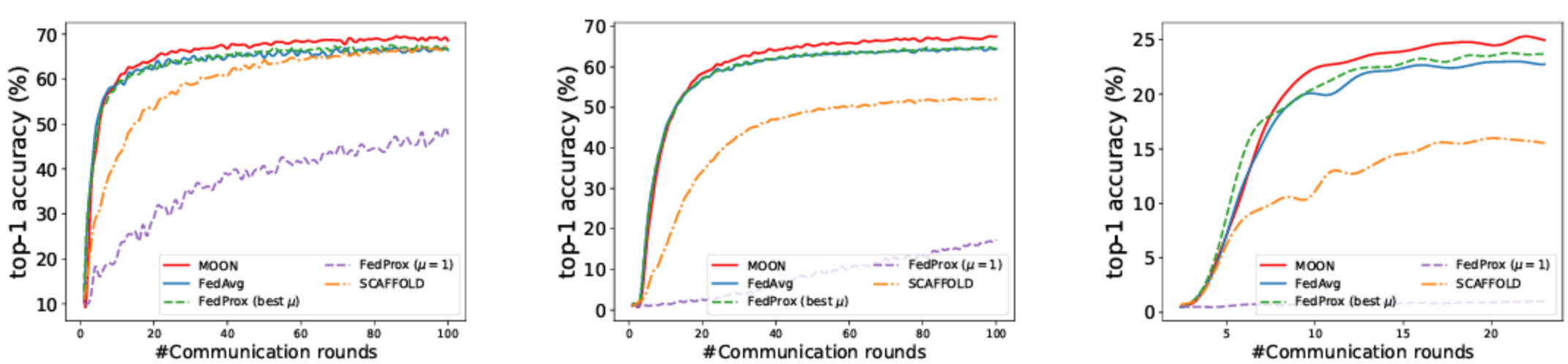


(c) Tiny-Imagenet

Method	CIFAR-10	CIFAR-100	Tiny-Imagenet
MOON	**69.1%**±0.4%	**67.5%**±0.4%	**25.1%**±0.1%
FedAvg	66.3%±0.5%	64.5% ±0.4%	23.0%±0.1%
FedProx	66.9%±0.2%	64.6%±0.2%	23.2%±0.2%
SCAFFOLD	66.6%±0.2%	52.5% ±0.3%	16.0%±0.2%
SOLO	46.3% ±5.1%	22.3%±1.0%	8.6%±0.4%

The top-1 accuracy of MOON and the other baselines on test datasets

Communication Efficiency



(a) CIFAR-10 (b) CIFAR-100 (c) Tiny-Imagenet

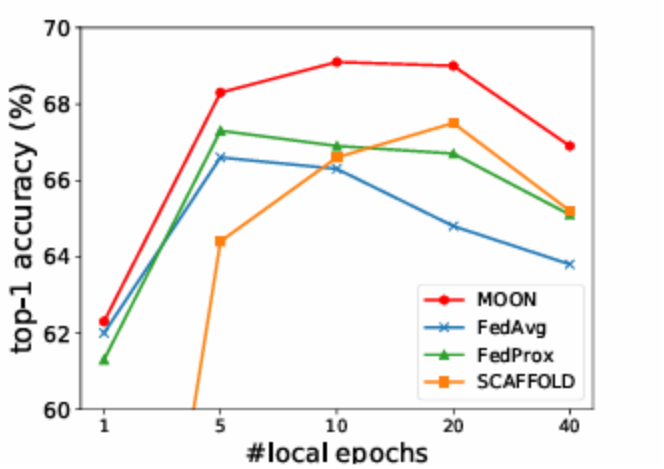
Figure 6. The top-1 test accuracy in different number of communication rounds. For FedProx, we report both the accuracy with best μ and the accuracy with $\mu = 1$.

Method	CIFAR-10		CIFAR-100		Tiny-Imagenet	
	#rounds	speedup	#rounds	speedup	#rounds	speedup
FedAvg	100	1×	100	1×	20	1×
FedProx	52	1.9×	75	1.3×	17	1.2×
SCAFFOLD	80	1.3×	\	<1×	\	<1×
MOON	**27**	**3.7**×	**43**	**2.3**×	**11**	**1.8**×

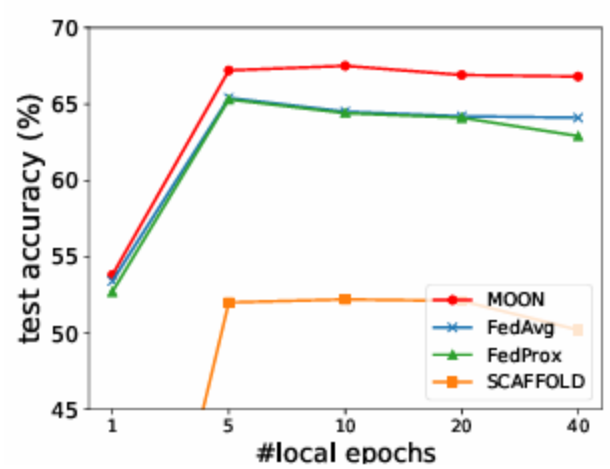
The number of rounds of different approaches to achieve the same accuracy as running FedAvg for 100 rounds

Number of Local Epochs

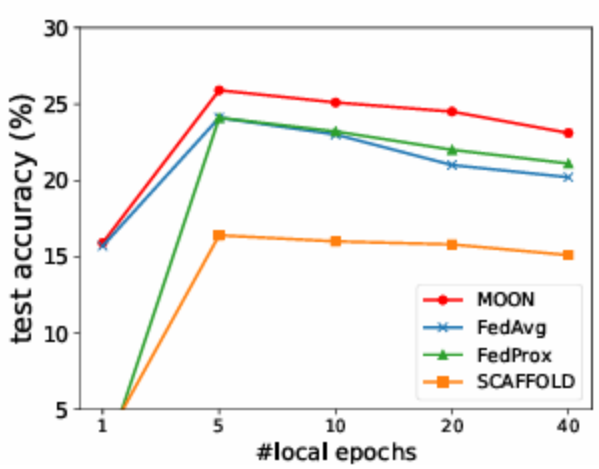
We study the effect of number of local epochs on the accuracy of final model.
When the number of local epochs is 1, the local update is very small. Thus, the training is slow and the accuracy is relatively low given the same number of communication rounds.
When the numberof local epochs becomes too large, the accuracy of all approaches drops, which is due to the drift of local updates, i.e., the local optima are not consistent with the global optima.



(a) CIFAR-10



(b) CIFAR-100



(c) Tiny-Imagenet

Scalability

To show the scalability of MOON, we try a larger number of parties on CIFAR-100. Specifically, we try two settings:

(1) We partition the dataset into 50 parties and all parties participate in federated learning in each round.

(2) We partition the dataset into 100 parties and randomly sample 20 parties to participate in federated learning in each round.

Method	#parties=50		#parties=100	
	100 rounds	200 rounds	250 rounds	500 rounds
MOON (μ=1)	54.7%	58.8%	54.5%	58.2%
MOON (μ=10)	**58.2%**	**63.2%**	**56.9%**	**61.8%**
FedAvg	51.9%	56.4%	51.0%	55.0%
FedProx	52.7%	56.6%	51.3%	54.6%
SCAFFOLD	35.8%	44.9%	37.4%	44.5%
SOLO	10%±0.9%		7.3%±0.6%	

Loss Function

To maximize the agreement between the representation learned by the global model and the representation learned by the local model, our model-contrastive loss ℓcon is proposed inspired by NT-Xent loss. Another intuitive option is to use ℓ2 regularization, and the local loss is

$$\ell = \ell_{sup} + \mu \left\| z - z_{glob} \right\|_2$$

We can observe that simply using ℓ2 norm even cannot improve the accuracy compared with FedAvg on CIFAR-10. While using ℓ2 norm can improve the accuracy on CIFAR-100 and Tiny-Imagenet, the accuracy is still lower than MOON.

Table 5. The top-1 accuracy with different kinds of loss for the second term of local objective. We tune μ from $\{0.001, 0.01, 0.1, 1, 5, 10\}$ for the ℓ_2 norm approach and report the best accuracy.

second term	CIFAR-10	CIFAR-100	Tiny-Imagenet
none (FedAvg)	66.3%	64.5%	23.0%
ℓ_2 norm	65.8%	66.9%	24.0%
MOON	**69.1%**	**67.5%**	**25.1%**

Conclusion

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

To improve the performance of federated deep learning models on non-IID datasets, we propose model-contrastive learning (MOON), a simple and effective approach for federated learning. MOON introduces a new learning concept, i.e., contrastive learning in model-level. Our extensive experiments show that MOON achieves significant improvement over state-of-the-art approaches on various image classification tasks. As MOON does not require the inputs to be images, it potentially can be applied to non-vision problems.

Thanks