Model-Contrastive Federated Learning

Qinbin Li[1], Bingsheng He[1], and Dawn Song[2]
[1]National University of Singapore, [2]UC Berkeley

Presented by Rui Chen

CVPR 2021

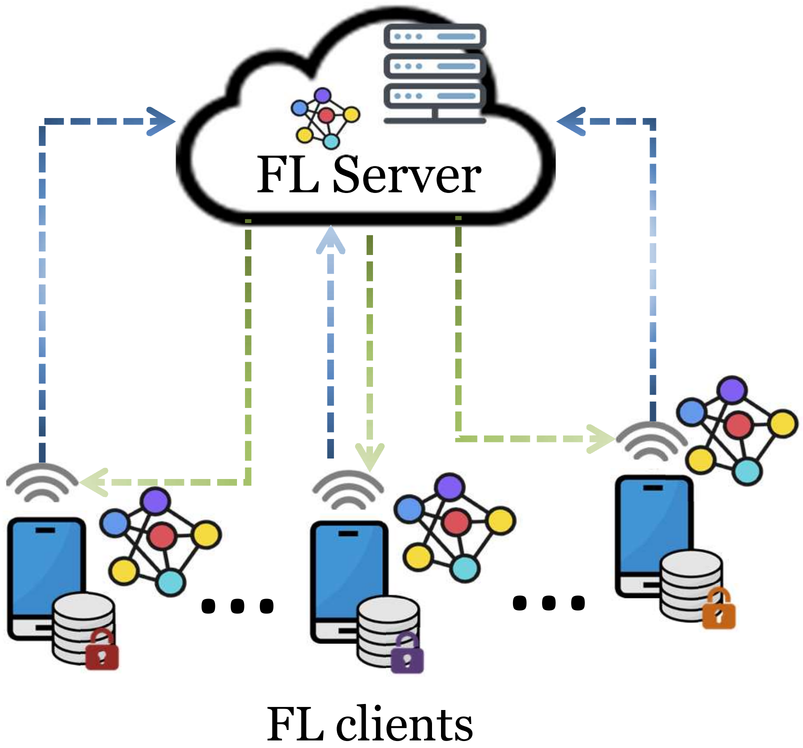
Outline

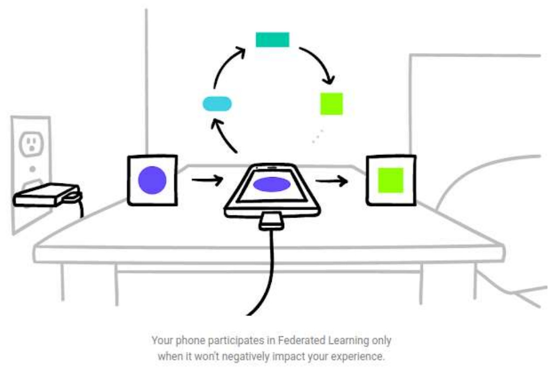
- Motivation
- Model-Contrastive Federated Learning
- Experimental Results
- Conclusion

ANTS LAB

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Motivation: Federated Learning





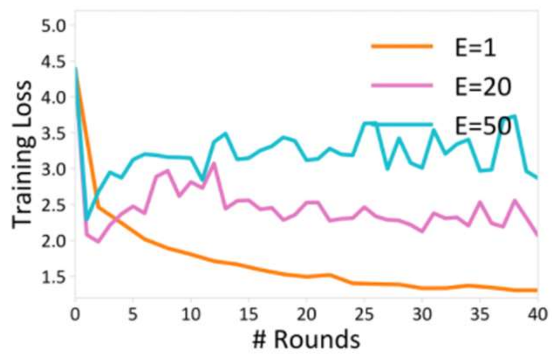
- Data never leaves local devices
- Learn on fresh real-world data

Heterogeneity in FL

statistical heterogeneity

highly non-identically distributed data

too much local work can hurt convergence



systems heterogeneity

stragglers

dropping slow devices can exacerbate convergence issues

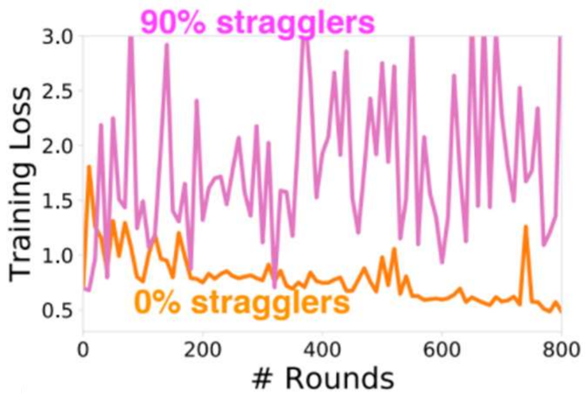
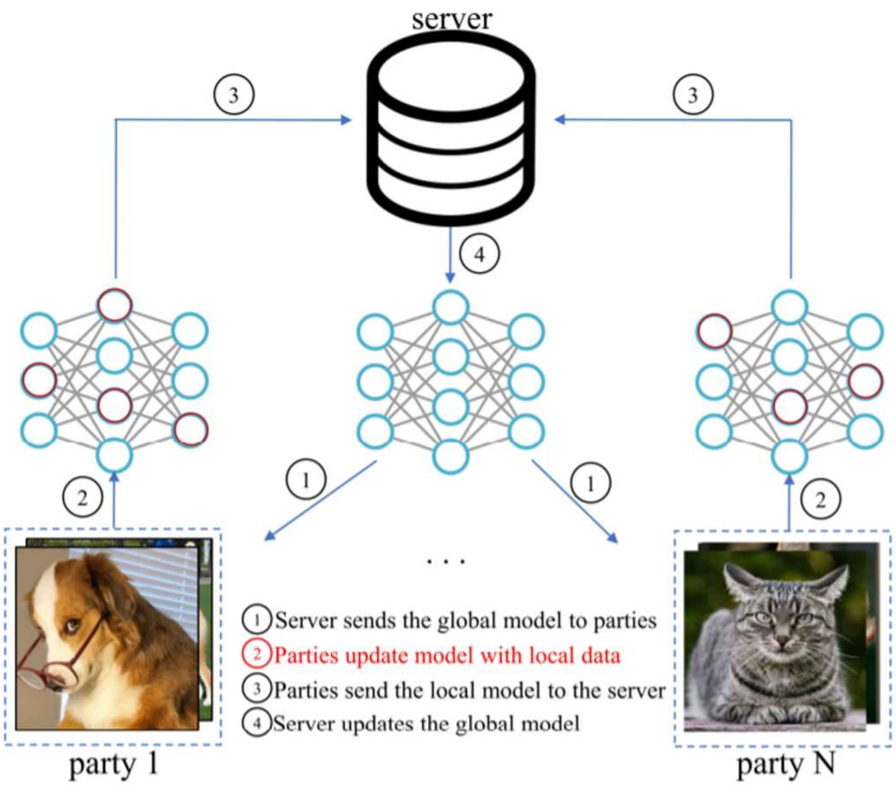


Figure from "Learning in Heterogeneous Networks: Optimization and Fairness"

Tackling Non-IID Data



- Local update stage ②

  FedProx: directly limits the local updates by ℓ_2-norm distance

  SCAFFOLD: corrects the local updates via variance reduction

  Those studies have little or even no advantage over FedAvg when deep models are used for training.

- Aggregation stage ④

  FedMA: match and average weights in a layer-wise manner.

  FedAvgM: applies momentum for global model updates

  FedNova: normalizes the local updates before averaging

ANTS LAB

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Contribution

- Address the non-IID issue from a novel perspective, i.e., model representation in the local update stage
- Propose model-contrastive learning (MOON), which conducts contrastive learning in model-level by comparing the representations learned by different models.

Outline

- Motivation
- Model-Contrastive Federated Learning
- Experimental Results
- Conclusion

Model-Contrastive Federated Learning

Contrastive learning –SimCLR [1]

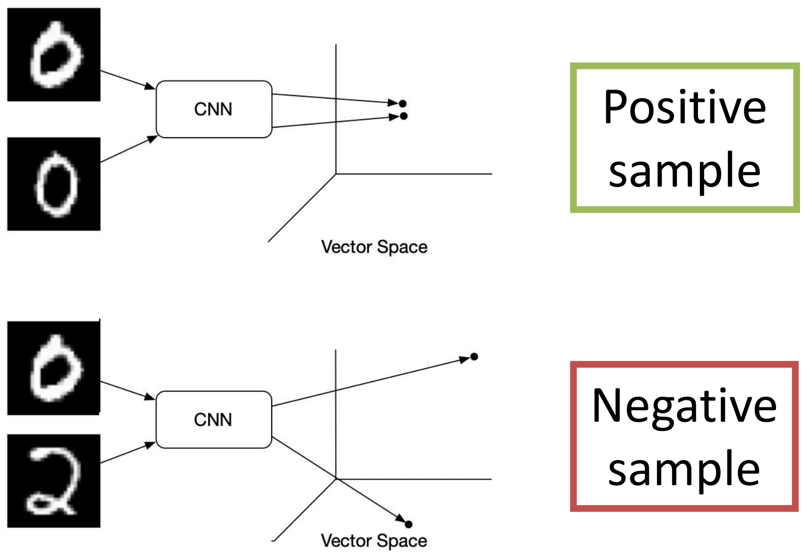
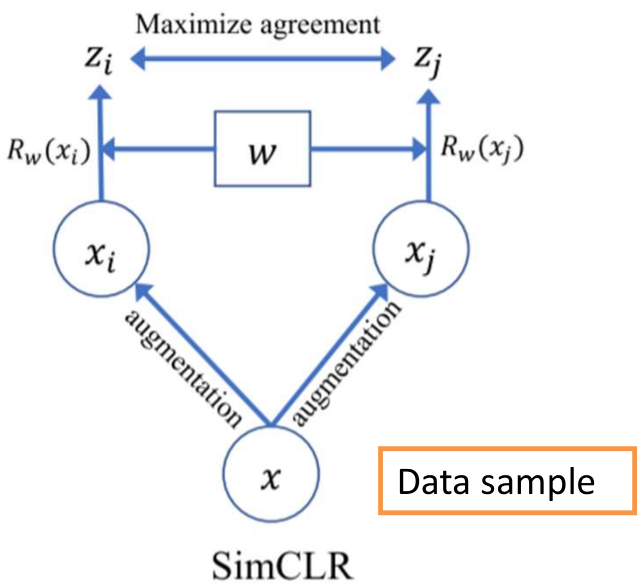


Figure from "Contrastive Loss Explained"

unsupervised learning



Here x denotes an image, w denotes a model, and R denotes the function to compute representation.

[1] T. Chen, et al. A simple framework for contrastive learning of visual representations. ICML 2020

Model-Contrastive Federated Learning

Observation: the global model trained on a *whole dataset* is able to learn a better representation than the local model trained on a *skewed subset*.

(a) global model

(b) local model

cannot be distinguished

Figure 2. T-SNE visualizations of hidden vectors on CIFAR-10.

Model-Contrastive Federated Learning

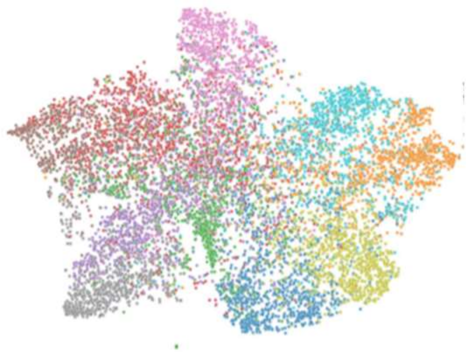
FedAvg global model can learn better than the local model.

The local training phase even leads the model to learn a worse representation.

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

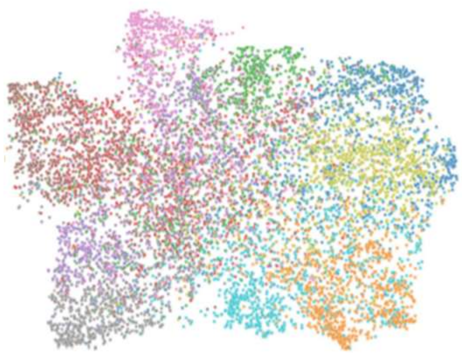
Model-Contrastive Federated Learning

MOON aims to decrease the distance between the representation learned by the local model and the representation learned by the global model



(c) FedAvg global model

Positive model

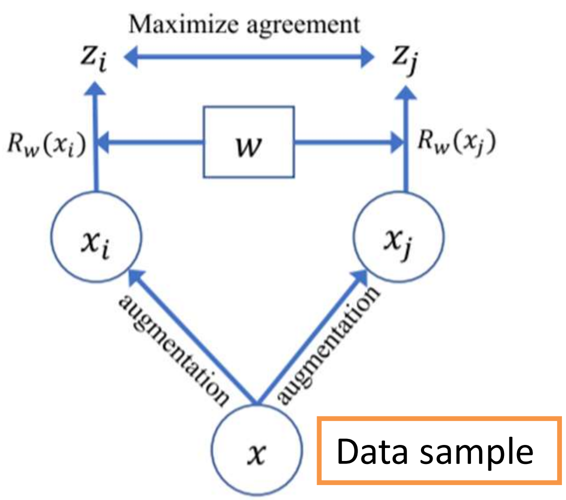


(d) FedAvg local model

Negative model

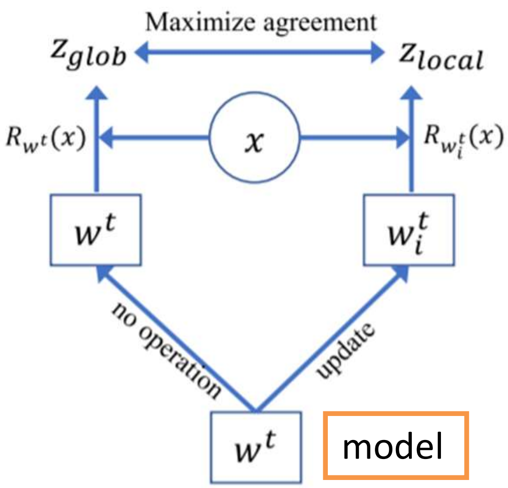
Model-Contrastive Federated Learning

unsupervised learning



SimCLR

supervised learning

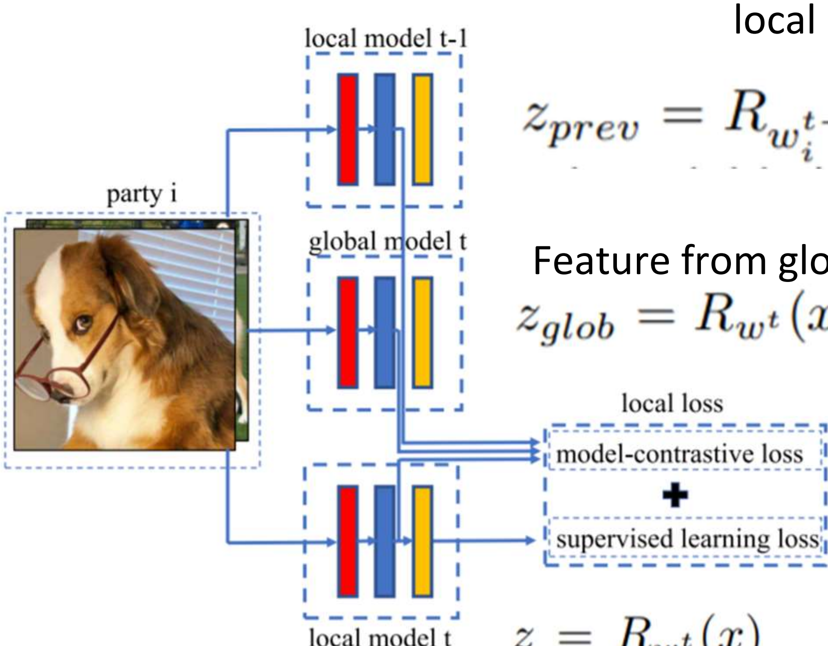


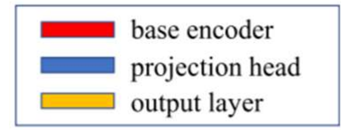
MOON

Here x denotes an image, w denotes a model, and R denotes the function to compute representation.

Model-Contrastive Federated Learning

Local training phase





Feature from last local model

$$z_{prev} = R_{w_i^{t-1}}(x)$$

Feature from global model

$$z_{glob} = R_{w^t}(x)$$

$$z = R_{w_i^t}(x)$$

Contrastive loss

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

Novel Local Loss

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

- Our approach is robust regardless of different amount of drifts.

Model-Contrastive Federated Learning

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

    $w_i^t \leftarrow$ $\boxed{\textbf{PartyLocalTraining}(i, w^t)}$

  $w^{t+1} \leftarrow \sum_{k=1}^{N} \frac{|\mathcal{D}^i|}{|\mathcal{D}|} w_k^t$

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**

    $\ell_{sup} \leftarrow CrossEntropyLoss(F_{w_i^t}(x), y)$

    $z \leftarrow R_{w_i^t}(x)$

    $z_{glob} \leftarrow R_{w^t}(x)$

    $z_{prev} \leftarrow R_{w_i^{t-1}}(x)$

    $\ell_{con} \leftarrow$

    $-\log \frac{\exp(\mathrm{sim}(z, z_{glob})/\tau)}{\exp(\mathrm{sim}(z, z_{glob})/\tau) + \exp(\mathrm{sim}(z, z_{prev})/\tau)}$

    $\ell \leftarrow \ell_{sup} + \mu \ell_{con}$

    $w_i^t \leftarrow w_i^t - \eta \nabla \ell$

return w_i^t to server

Outline

- Motivation
- Model-Contrastive Federated Learning
- Experimental Results
- Conclusion

Experiment Setup

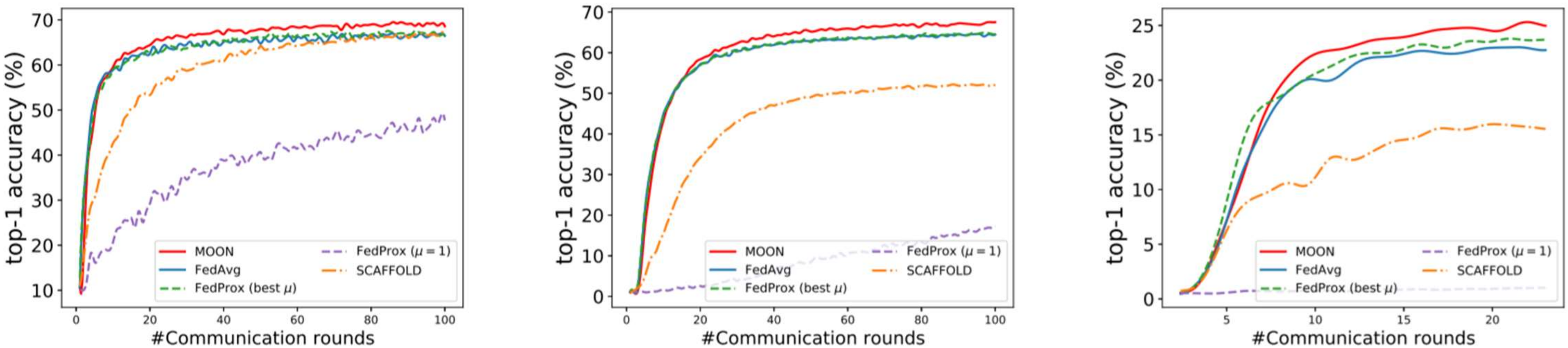
- We use PyTorch to implement MOON and the other baselines
- Dataset: CIFAR10, CIFAR100, tiny-ImageNet
- Encoder: CNN for CIFAR10; ResNet-50 for CIFAR100 & tiny-ImageNet
- Data partition: apply Dirichlet distribution (Dir(β)) to generate the non-IID data partition among parties.
- A 2- layer MLP is used as the projection head.
- Baseline: FedAvg ; FedProx ; SCAFFOLD and Local training

Performance evaluation

Table 1. The top-1 accuracy of MOON and the other baselines on test datasets. For MOON, FedAvg, FedProx, and SCAFFOLD, we run three trials and report the mean and standard derivation. For SOLO, we report the mean and standard derivation among all parties.

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%

Performance evaluation



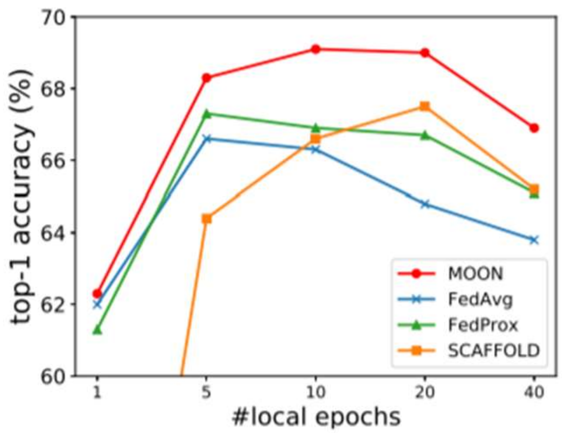
(a) CIFAR-10 (b) CIFAR-100 (c) Tiny-Imagenet

- MOON is much more communication-efficient than the other approaches.

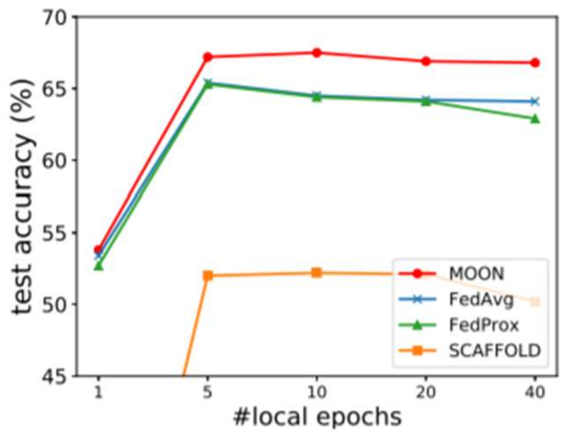
Table 2. The number of rounds of different approaches to achieve the same accuracy as running FedAvg for 100 rounds (CIFAR-10/100) or 20 rounds (Tiny-Imagenet). The speedup of an approach is computed against FedAvg.

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**×

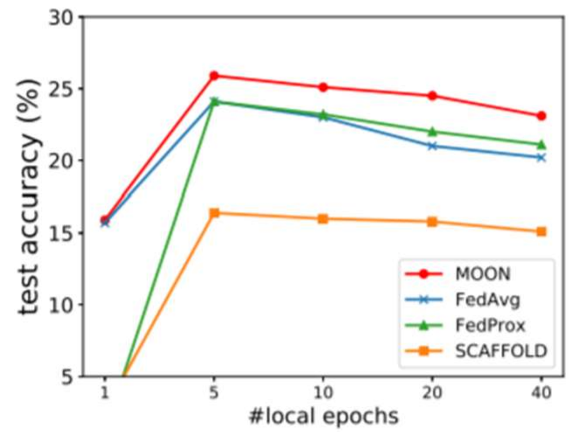
Performance evaluation



(a) CIFAR-10



(b) CIFAR-100



(c) Tiny-Imagenet

- MOON can effectively mitigate the negative effects of the drift by too many local updates.

ANTS LAB

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Performance evaluation

- Scalability

Table 3. The accuracy with 50 parties and 100 parties (sample fraction=0.2) on CIFAR-100.

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%	

- Diversity

Table 4. The test accuracy with β from $\{0.1, 0.5, 5\}$.

Method	$\beta = 0.1$	$\beta = 0.5$	$\beta = 5$
MOON	**64.0%**	**67.5%**	**68.0%**
FedAvg	62.5%	64.5%	65.7%
FedProx	62.9%	64.6%	64.9%
SCAFFOLD	47.3%	52.5%	55.0%
SOLO	15.9%±1.5%	22.3%±1%	26.6%±1.4%

- MOON can outperform the other approaches a lot with more communication rounds.

Performance evaluation

- The average training time per round

Table 11. The average training time per round.

Method	CIFAR-10	CIFAR-100	Tiny-Imagenet
FedAvg	330s	20min	103min
FedProx	340s	24min	135min
SCAFFOLD	332s	20min	112min
MOON	337s	31min	197min

Conclusion

- MOON is a simple and effective federated learning framework.
- MOON addresses the non-IID data issue with the novel design of model-based contrastive learning.


THANK YOU
UNIVERSITY of HOUSTON | ENGINEERING

Performance evaluation

- Selection of different local loss function

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%**