Supplementary Material: 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

1. More Details of the Datasets

The statistics of each dataset are shown in Table 1 when setting $\beta = 0.5$. All datasets provide a training dataset and test dataset. All the reported accuracies are computed on the test dataset.

Table 1. The statistics of datasets.

dataset	#training samples/party		#test samples
	mean	std	
CIFAR10	5,000	1,165	10,000
CIFAR100	5,000	181	10,000
Tiny-Imagenet	10,000	99	10,000

Figure 1 and Figure 2 show the data distribution of $\beta = 0.1$ and $\beta = 5$ (used in Section 4.6 of the main paper), respectively.

2. Projection Head

We use a projection head to map the representation like [1]. Here we study the effect of the projection head. We remove the projection head and conduct experiments on CIFAR-10 and CIFAR-100 (Note that the network architecture changes for all approaches). The results are shown in Table 2. We can observe that MOON can benefit a lot from the projection head. The accuracy of MOON can be improved by about 2% on average with a projection head.

Table 2. The top1-accuracy with/without projection head.

Method		CIFAR-10	CIFAR-100
without projection head	MOON	66.8%	66.1%
	FedAvg	66.7%	65.0%
	FedProx	67.5%	65.4%
	SCAFFOLD	67.1%	49.5%
	SOLO	39.8%±3.9%	22.5%±1.1%
with projection head	MOON	**69.1%**	**67.5%**
	FedAvg	66.3%	64.5%
	FedProx	66.9%	64.6%
	SCAFFOLD	66.6%	52.5%
	SOLO	46.3%±5.1%	22.3%±1.0%

3. IID Partition

To further show the effect of our model-contrastive loss, we compare MOON and FedAvg when there is no heterogeneity among local datasets. The dataset is randomly and equally partitioned into the parties. The results are shown in Table 3. We can observe that the model-contrastive loss has little influence on the training when the local datasets are IID. The accuracy of MOON is still very close to FedAvg even though with a large μ. MOON is still applicable when there is no heterogeneity issue in data distributions across parties.

Table 3. The top-1 accuracy of MOON and FedAvg with IID data partition on CIFAR-10.

Method		Top-1 accuracy
MOON	$\mu = 0.1$	73.6%
	$\mu = 1$	73.6%
	$\mu = 5$	73.0%
	$\mu = 10$	72.8%
FedAvg ($\mu = 0$)		73.4%

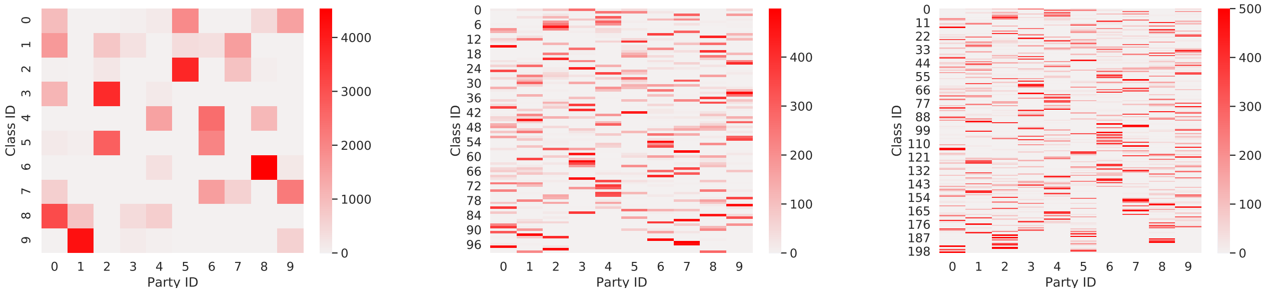
4. Hyper-Parameters Study

4.1. Effect of μ

We show the accuracy of MOON with different μ in Table 4. The best μ for CIFAR-10, CIFAR-100, and Tiny-Imagenet are 5, 1, and 1, respectively. When μ is set to a small value (i.e., $\mu = 0.1$), the accuracy of MOON is very close to FedAvg (i.e., $\mu = 0$) since the impact of model-contrastive loss is small. As long as we set $\mu \geq 1$, MOON can benefit a lot from the model-contrastive loss. Overall, we find that $\mu = 1$ is a reasonable good choice if they do not want to tune the parameter, where MOON achieves at least 2% higher accuracy than FedAvg.

4.2. Effect of temperature and output dimension

We tune τ from $\{0.1, 0.5, 1.0\}$ and tune the output dimension of projection head from $\{64, 128, 256\}$. The results are shown in Figure 3. The best τ for CIFAR-10, CIFAR-100, and Tiny-Imagenet are 0.5, 1.0, and 0.5,

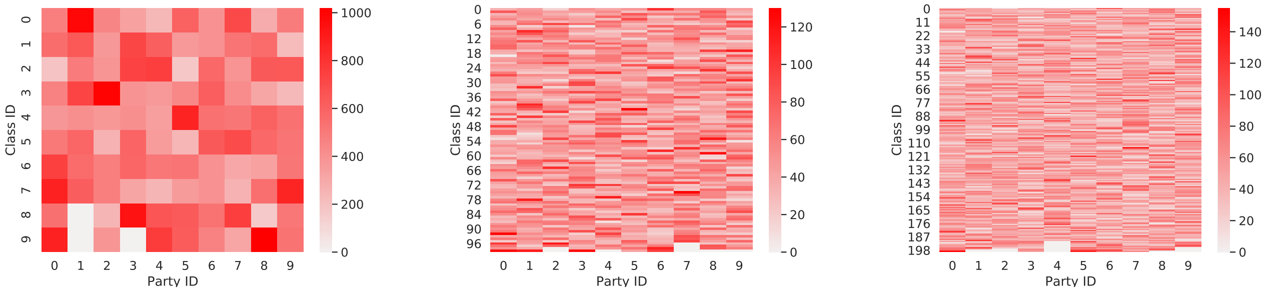


(a) CIFAR-10

(b) CIFAR-100

(c) Tiny-Imagenet

Figure 1. The data distribution of each party using non-IID data partition with $\beta = 0.1$.



(a) CIFAR-10

(b) CIFAR-100

(c) Tiny-Imagenet

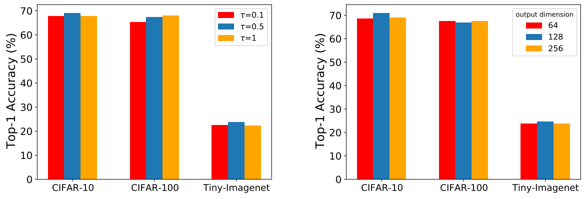
Figure 2. The data distribution of each party using non-IID data partition with $\beta = 5$.

Table 4. The test accuracy of MOON with μ from {0, 0.1, 1, 5, 10}. Note that MOON is actually FedAvg when $\mu = 0$.

μ	CIFAR-10	CIFAR-100	Tiny-Imagenet
0	66.3%	64.5%	23.0%
0.1	66.5%	65.1%	23.4%
1	68.4%	**67.5%**	**25.1%**
5	**69.1%**	67.1%	24.4%
10	68.3%	67.3%	25.0%

respectively. The best output dimension for CIFAR-10, CIFAR-100, and Tiny-Imagenet are 128, 256, and 128, respectively. Generally, MOON is stable regarding the change of temperature and output dimension. As we have shown in the main paper, MOON already improves FedAvg a lot with a default setting of temperature (i.e., 0.5) and output dimension (i.e., 256). Users may tune these two hyper-parameters to achieve even better accuracy.

5. Combining with FedAvgM



(a) The effect of τ

(b) The effect of output dimension

Figure 3. The top-1 accuracy of MOON trained with different temperatures and output dimensions.

Table 5. The combining of MOON and FedAvgM.

Datasets	FedAvg	MOON	FedAvgM	MOON+FedAvgM
CIFAR-10	66.3%	69.1%	67.1%	69.6%
CIFAR-100	64.5%	67.5%	65.1%	67.8%
Tiny-Imagenet	23.0%	25.1%	23.4%	25.5%

As we have mentioned in the fourth paragraph of Section 2.1, MOON can be combined with the approaches working on improving the aggregation phase. Here we combine MOON and FedAvgM [2]. We tune the server momentum $\beta \in \{0.1, 0.7, 0.9\}$. With the default experimental setting in Section 4.1, the results are shown in Table 5. While FedAvgM is better than FedAvg, MOON can further improve FedAvgM by 2-3%.

6. Computation Cost

Since MOON introduces an additional loss term in the local training phase, the training of MOON will be slower than FedAvg. For the experiments in Table 1, the average

Table 6. 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

Table 7. The effect of maximum number of negative pairs. We tune μ from $\{0.1, 1, 5, 10\}$ for all approaches and report the best accuracy.

maximum number of negative pairs	top-1 accuracy
$k = 1$	**69.1%**
$k = 2$	67.2%
$k = 5$	67.7%
$k = 100$	63.5%

training time per round with a NVIDIA Tesla V100 GPU and four Intel Xeon E5-2684 20-core CPUs are shown in Table 6. Compared with FedAvg, the computation overhead of MOON is acceptable especially on CIFAR-10 and CIFAR-100.

7. Number of Negative Pairs

In typical contrastive learning, the performance usually can be improved by increasing the number of negative pairs (i.e., views of different images). In MOON, the negative pair is the local model being updated and the local model from the previous round. We consider using a single negative pair during training in the main paper. It is possible to consider multiple negative pairs if we include multiple local models from the previous rounds. Suppose the current round is t. Let k denotes the maximum number of negative pairs. Let $z^i_{prev} = R_{w_i^{t-i}}(x)$ (i.e., z^i_{prev} is the representation learned by the local model from $(t-i)$ round). Then, our local objective is

$$\ell_{con} = -\log \frac{\exp(\text{sim}(z, z_{glob})/\tau)}{\exp(\text{sim}(z, z_{glob})/\tau) + \sum_{i=1}^{k} \exp(\text{sim}(z, z^i_{prev})/\tau)} \tag{1}$$

If $k = 1$, then the objective is the same as MOON presented in the main paper. If $k > t$, since there are at most t local models from previous rounds, we only consider the previous t local models (i.e., only t negative pairs). There is no model-contrastive loss if $t = 0$ (i.e., the first round). Here we study the effect of the maximum number of negative pairs on CIFAR-10. The results are shown in Table 7. Unlike typical contrastive learning, the accuracy of MOON cannot be increased by increasing the number of negative pairs. MOON can achieve the best accuracy when $k = 1$, which is presented in our main paper.

References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[2] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.