Supplementary for *Collaborative Unsupervised Visual Representation Learning from Decentralized Data*

Weiming Zhuang[1,3] Xin Gan[2] Yonggang Wen[2] Shuai Zhang[3] Shuai Yi[3]
[1]S-Lab, Nanyang Technological University, [2]Nanyang Technological University, [3]SenseTime Research
{weiming001,ganx0005}@e.ntu.edu.sg,ygwen@ntu.edu.sg,{zhangshuai,yishuai}@sensetime.com

Abstract

This supplementary material provides more implementation details and presents more experimental results, including ablation studies on DAPU, batch size, and the number of clients. We also present the convergence of FedU and more t-SNE visualization of representations of different settings.

1. Implementation Details

In this section, we provide more implementation details on image augmentations, semi-supervised evaluation, and transfer learning evaluation.

Image Augmentations We adopt the image augmentations from BYOL [3] and SimCLR [1]. We select a random patch of the image and resize it to 32x32 for CIFAR datasets. After that, two transformations are applied to the image: a random horizontal flip and a color distortion.

Semi-supervised Learning In semi-supervised evaluation protocol, we train models with only unlabeled data — excluding the 1% or 10% labeled data. These 1% and 10% labeled data are only used in fine-tuning the trained models with an additional classifier.

Transfer Learning In transfer learning evaluation protocol, we train models using Mini-ImageNet [5] dataset and fine-tune on CIFAR [4] datasets. Images in Mini-ImageNet have size 84x84, while images in CIFAR datasets are 32x32. We scale image size of CIFAR datasets to be 84x84 in fine-tuning.

2. Experimental Results

Communication Protocol Table 1 shows that aggregating and uploading the online encoder achieves the best performance. We run these experiments with ResNet-18 on CIFAR-100 non-IID setting. It complements the results of Table 4 in the main manuscript.

Divergence-aware Predictor Update Table 2 shows that DAPU outperforms always updating the predictors of clients using either the local or global predictor on both IID and non-IID settings. It complements the results of Figure 5(a) in the main manuscript, which only presents that DAPU outperforms other methods on the non-IID setting. Besides, Figure 2 presents the impact of threshold μ on the non-IID setting of the CIFAR-100 dataset, complementing the results on the CIFAR-10 dataset in the main manuscript (Figure 5(b)). It also indicates that $\mu = 0.2$ achieves the best performance. We run these experiments with $E = 1$ and $R = 800$.

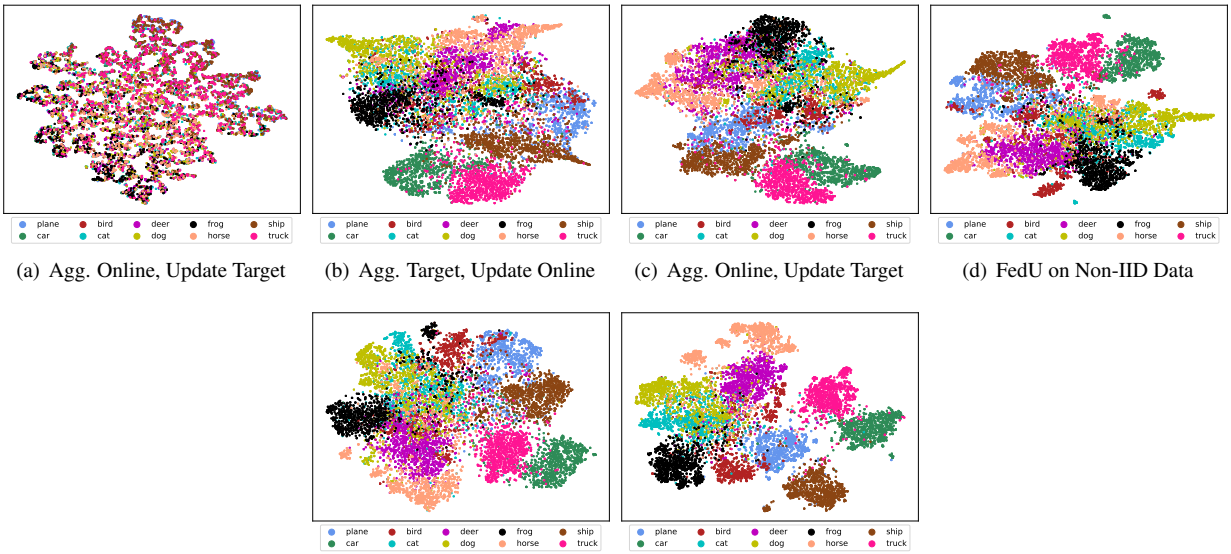
Aggregate	Update	Accuracy (%)
Online	**Online**	**62.86**
Online	Target	3.44
Online	Both	60.78
Target	Online	48.10
Target	Target	10.27
Target	Both	61.41

Table 1. Top-1 accuracy comparison of using the *online encoder* or *target encoder* for aggregation and update. Both means updating both encoders. Aggregating and updating the online encoder achieves the best performance.

Update Method	CIFAR-10		CIFAR-100	
	Non-IID	IID	Non-IID	IID
Local Pred.	82.18	91.29	61.69	67.41
Global Pred.	84.07	91.41	63.30	67.56
DAPU	**87.14**	**93.13**	**68.02**	**67.66**

Table 2. Top-1 accuracy comparison of DAPU and always updating with the local or global predictor. DAPU outperforms other methods in all settings.

Impact of Batch Size We study the impact of batch size in Table 3. Constant learning rate (LR) means that we use the same learning rate $\eta = 0.032$ for experiments of different batch sizes. As for adjusted LR, we use learning rate $\eta = \frac{B \times 0.032}{128}$ for different batch sizes B. By adjusting the learning rate accordingly, the results are similar among batch sizes $B = \{32, 64, 128, 256\}$. However, when the learning rate is constant, a larger batch size leads



(a) Agg. Online, Update Target (b) Agg. Target, Update Online (c) Agg. Online, Update Target (d) FedU on Non-IID Data

(e) FedU on IID Data (f) BYOL [3] (Centralized)

Figure 1. T-SNE visualization of representations learned from different methods: (a) Aggregate the online encoder and update the target encoder; (b) Aggregate the target encoder and update the online encoder; (c) Aggregate and update the online encoder; (d) Our proposed FedU trained on non-IID CIFAR-10 data; (e) FedU trained on IID data; (f) Centralized unsupervised learning method (BYOL [3]). (a), (b), and (c) always use the global predictor, while (d) uses DAPU to dynamically update the predictor. FedU with DAPU (d) presents better clustering results than (a), (b), and (c). FedU's representation learned from IID data (e) is also comparable with centralized training (f).

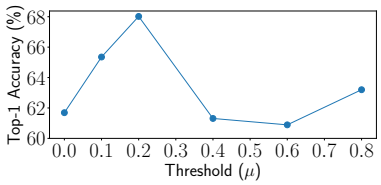


Figure 2. Impact of threshold μ on the non-IID setting of the CIFAR-100 dataset. DAPU with $\mu = 0.2$ achieves the best performance, complementing our results in the main manuscript.

Batch Size	Adjusted LR	Constant LR
32	81.03	74.92
64	81.69	79.59
128	81.70	81.70
256	81.80	83.22

Table 3. Impact of batch size on performance. A larger batch size leads to better performance when the learning rate is constant. The performances of various batch sizes are similar when the learning rate is adjusted accordingly.

to better performance. We use $B = 128$ in our main manuscript. It indicates that the experimental results in the main manuscript can be further improved with larger batch size. We run these experiments with $E = 5$ and $R = 100$ on non-IID setting of CIFAR-10.

Convergence of FedU Figure 3 shows that FedU has nice convergence property — the accuracy steadily improves as training proceeds. We monitor the training progress by performing classification using k-nearest neighbors (kNN) [6, 2]. We set the number of neighbors to 200 and the temperature to 0.1. These experiments are run with $E = 5$ and $R = 100$.

Scalability of FedU We compare the performance of different numbers of clients K in Figure 4. We use IID setting to conduct the experiments to keep the same data distribution as we change K among $\{1, 2, 5, 8, 10\}$. We split the CIFAR-10 dataset to K clients with equal data volume, clients of larger K contain less data. Although the performance decreases with the increase of K, it is still better than single client training when $K = 10$. Besides, the performance almost maintains from $K = 5$ to $K = 10$. We run these experiments with $E = 5$ and $R = 100$.

Representation Comparison We compare the t-SNE visualization of representations learned from different methods in Figure 1. Figure 1(a), 1(b), and 1(c) always uses the global predictor to update clients' predictors, complementing the t-SNE visualizations on the main manuscript where the local predictor are always used. FedU with DAPU 1(d) has better clustering results than the first three, indicating the effectiveness of DAPU. Besides, FedU's representation learned from IID data (Figure 1(e)) is also comparable with centralized training (Figure 1(f)).

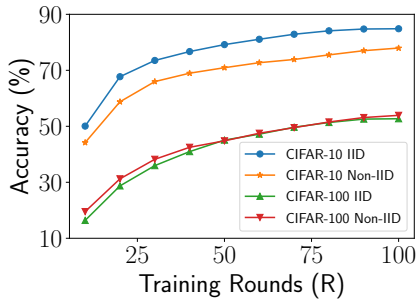


Figure 3. The kNN testing accuracy improves as training continues, demonstrating the convergence of FedU.

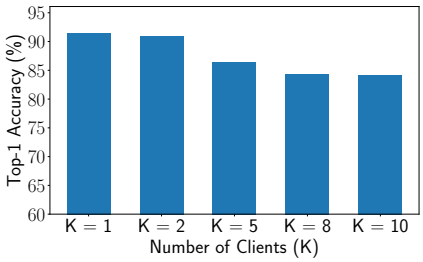


Figure 4. Impact of the number of clients on FedU. Although the performance decreases as K increases, the performance of $K = 10$ is still better than single client training.

References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1

[2] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *arXiv preprint arXiv:2011.10566*, 2020. 2

[3] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc., 2020. 1, 2

[4] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1

[5] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. *arXiv preprint arXiv:1606.04080*, 2016. 1

[6] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018. 2