L-DAWA: Layer-wise Divergence Aware Weight Aggregation in Federated Self-Supervised Visual Representation Learning (Supplementary)

Yasar Abbas Ur Rehman[1,*], Yan Gao[2,*], Pedro Porto Buarque de Gusmão[2], Mina Alibeigi[2,3], Jiajun Shen[1], Nicholas D. Lane[2,4]

[1]TCL AI Lab, Hong Kong, [2]University of Cambridge, United Kingdom, [3]Zenseact, Sweden, [4]Flower Labs

1. Appendix

1.1. SimCLR vs. Barlow Twins loss function

$$L_{SimCLR} = u^T v^+/\tau - log \sum_{v \in \{v^+, v^-\}} exp(u^T v/\tau) \quad (1)$$

The above equation represents the NT-Xent (Normalized Temperature-scaled Cross Entropy) loss function as proposed in [1]. The input u^T, v^+, and v^- are l_2 normalized. τ is a temperature coefficient.

$$L_{BarlowTwins} = \sum_i (1 - C_{ii})^2 + \lambda \sum_i \sum_{j \neq i} C_{ij}^2 \quad (2)$$

In the above equation, C represents the cross-correlation matrix computed on the output features of the two identical networks along the batch dimension, as illustrated in [7]. C_{ii} represents the diagonal elements of the cross-correlation matrix while C_{ij} represents the off-diagonal elements of the cross-correlation matrix. λ is a positive hyperparameter that controls the trade-off between the first term (invariance) and the second term (redundancy reduction).

1.2. Properties of L-DAWA

L-DAWA intrinsically provides significant performance improvement in *cross-silo* settings. Such performance improvement signifies the importance of introducing divergence control during weight aggregation. One can see from Table 1, that L-DAWA provides a lightweight aggregation method that is unbiased and independent of the metadata. L-DAWA provides layer-wise divergence control at the server, unlike FedU which provides partial divergence control for only the predictor network on the client side. One can further note from Table 1 that FedAvg, Loss, and FedU equally treat all the layers of the client's model by multiplying it with a constant coefficient. In contrast, L-DAWA treats each layer of the client's model by the measure of divergence that varies from layer to layer.

The current state-of-the-art aggregation methods (FedAvg, Loss, and FedU) can be improved by introducing a measure of the model quality based on *angular measure of divergence* as shown in Table 1. We note that individual bias toward sample size as in FedAvg and FedU, local loss in Loss aggregation strategies is effectively mitigated by introducing *angular measure of divergence* in these methods resulting in an improved and fair performance for both contrastive (SimCLR) and non-contrastive (Barlow Twins) SSL approaches.

1.3. Optimization trajectory analysis under loss landscape

We further explore the effects of L-DAWA on the model's global loss landscape and the global optimization trajectories under the *cross-silo* settings with SimCLR. For this purpose, we explore the loss landscape and optimization trajectories of FedAvg, Loss, and FedU compared with their divergence-controlled versions L-DAWA$_{FedAvg}$, L-DAWA$_{Loss}$ and L-DAWA$_{FedU}$. Interestingly, we find that the loss landscape of FedAvg, Loss, and FedU (Figure 1 (a-c)) is more chaotic and the global optimization trajectory fell into a narrow local minimum leading to sub-optimal performance. When L-DAWA is introduced into these aggregation methods, the global optimization trajectory ends up in a much wider basin of attraction in the loss landscape resulting in improved performance (Figure 1 (d-f)).

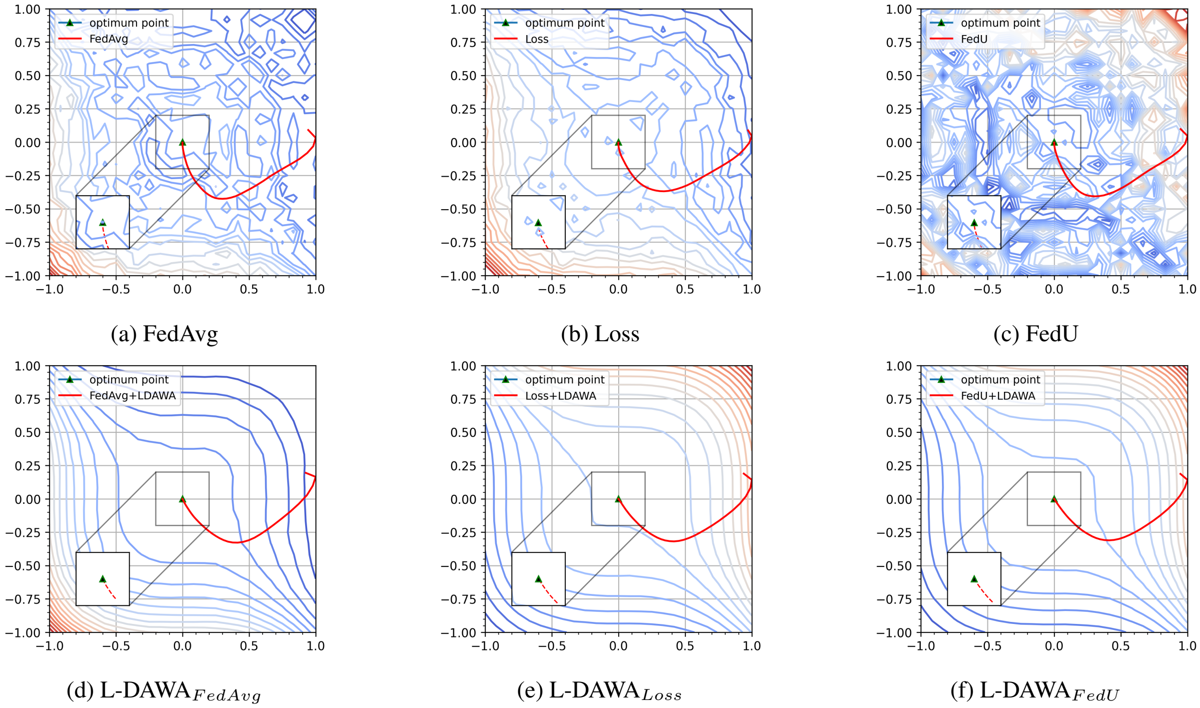
1.4. Divergence measurements over the clients

We find that L-DAWA reduces the angular divergence between the clients' models and the global model during FL pre-training by scaling each client with its measure of the *angular divergence* with respect to the global model. Such scaling controls the length of the step taken by the global model to reach the optimum point. For example, if the *angular divergence* between a certain client's model and the global model is higher, L-DAWA will downscale the contribution of such client's model based on the extent of the divergence. This results in the global model optimization trajectory being less affected by the diverging clients,

[*] Equal contribution, authors ordered alphabetically.

Method	Metadata Type	Bias	Div. Control	Weighting Coefficient		%Acc. Cross-Silo					
				Type	Nature	CIFAR-10		CIFAR-100		Tiny ImageNet	
						SimCLR	Barlow Twins	SimCLR	Barlow Twins	SimCLR	Barlow Twins
FedAvg	Sample Size	✓	None	Sample Prob.	Const.	71.07	65.02	43.85	35.70	32.92	15.40
Loss	Local Loss	✓	None	Local Loss	Const.	71.34	57.12	44.69	34.76	33.37	12.24
FedU	Sample Size	✓	Partial	Sample Prob.	Const.	70.36	64.55	44.31	35.25	32.63	15.16
L-DAWA	None	✗	Layer-wise	Layer-Wise Div.	Var.	75.60	69.31	49.88	41.85	37.22	21.47
L-DAWA$_{FedAvg}$	Sample Size	✓	Layer-wise	Layer-Wise Div. + Sample Prob.	Var.	75.72	**69.92**	49.99	41.49	36.97	21.58
L-DAWA$_{Loss}$	Local Loss	✓	Layer-wise	Layer-Wise Div. + Local Loss	Var.	**76.55**	69.46	50.29	**41.89**	37.12	11.90
L-DAWA$_{FedU}$	Sample Size	✓	Layer-wise	Layer-Wise Div. + Sample Prob.	Var.	76.23	69.50	**50.59**	41.72	**37.35**	**21.80**

Table 1: Comparison of properties and performances for L-DAWA with state-of-the-art aggregation methods. "Acc.", "Div.", "Prob.", "Const.", "Var." stands for Accuracy, Divergence, Probability, Constant, and Variable, respectively. Bias represents the deviation towards the meta data.



(a) FedAvg

(b) Loss

(c) FedU

(d) L-DAWA$_{FedAvg}$

(e) L-DAWA$_{Loss}$

(f) L-DAWA$_{FedU}$

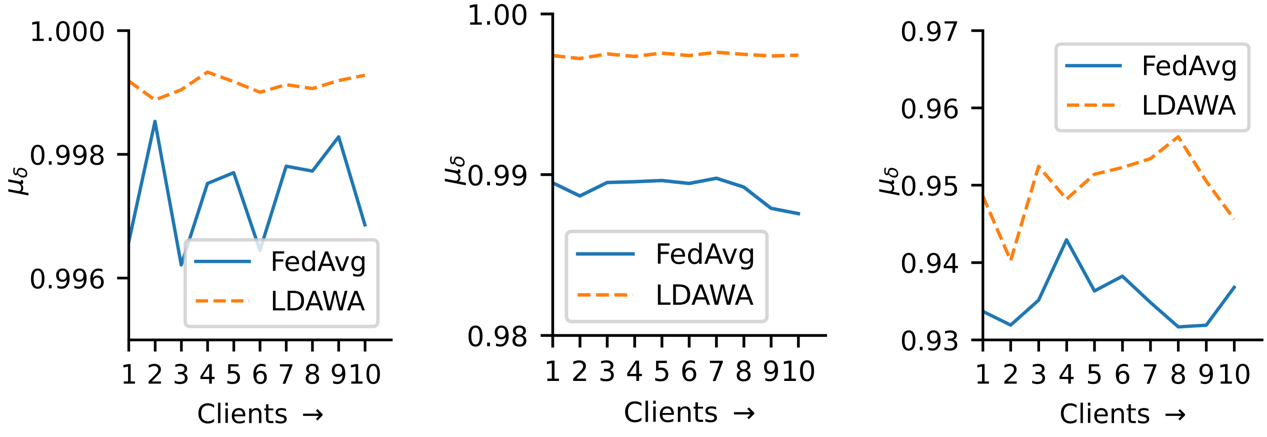
Figure 1: Illustration of global model's optimization trajectories and the loss landscape under various aggregation methods using SimCLR in FL settings.

which causes improved control over the divergence of individual clients' models with respect to the global model. To provide a formal illustration, we compute the average angular divergence for each client in *cross-silo* FL settings with SimCLR as shown in Figure 2. One can see from Figure 2 that the *angular divergence* (a.k.a. cosine of the angle between the global model and client's models) of FedAvg for all clients with increasing local epochs gets higher, resulting in lower values of mean angular divergence. On the other hand, L-DAWA maintains a steady angular divergence resulting in higher mean values for angular divergence.

1.5. Evaluation with individual clients fine-tuning

We empirically show that L-DAWA and L-DAWA combined with FedAvg (i.e., L-DAWA$_{FedAvg}$), provide nearly the same performance on all the participating clients in FL *cross-silo* settings. This phenomenon suggests that the inclusion of angular divergence measurement δ restricts the divergence bounds of the optimization trajectory in the global model. To simulate such an effect, we pre-trained a SimCLR on the Non-iid ($\alpha = 0.1$) version of CIFAR10 for $R = 200$ rounds under the *cross-silo* settings ($K = 10$) with FedAvg, L-DAWA, and L-DAWA$_{FedAvg}$. After pre-training, we fine-tune the last layer of the pre-trained global model on the individual clients' dataset that has participated in FL and subsequently evaluate it on the common CIFAR-10 test set (Table 2). From the results of Table 2, we note the following observations:

First, the clients with more data do not necessarily give better performance. One can see from Table 2 that the client



(a) E=1

(b) E=5

(c) E=10

Figure 2: The mean angular divergence between the clients' models with respect to the previous global model averaged over $R = 100$ rounds, computed by the following equation: $\mu_\delta = \frac{1}{R}\sum_{r=1}^{R}\delta_k^r$, where $k = \{1, ..., 10\}$. The higher δ value means lower divergence. L-DAWA has a good control of oscillations, maintaining the angular divergence in a lower level over FL rounds than FedAvg.

client →	1	2	3	4	5	6	7	8	9	10	Avg
Data Samples per client →	5839	163	2477	5706	8780	7183	8519	6518	2978	1802	4996.5
FedAvg	40.18	27.22	23.64	30.55	32.21	41.89	34.85	39.22	18.92	27.65	31.63
L-DAWA	42.57	29.97	26.95	32.64	33.68	45.02	36.74	43.58	19.69	29.93	**34.08**
L-DAWA$_{FedAvg}$	43.13	30.63	26.96	33.00	33.68	45.61	38.21	43.62	20.07	30.19	**34.51**

Table 2: Linar-probe accuracy of the FL (cross-silo) pre-trained model. After FL pre-training, we fine-tune the last layer of each client model with the client's local dataset and evaluate it on the CIFAR-10 test set. It can be seen that L-DAWA and L-DAWA$_{FedAvg}$ provide nearly the same test results suggesting that both L-DAWA and L-DAWA$_{FedAvg}$ may converge to the similar basin of loss landscape during FL pre-training.

$5, 7$ contain more than 8000 data samples, however, their performance on the test set is 32.21% and 34.85%, respectively. In contrast, client 1 contains 5839 samples while obtaining a much better performance of 40.18% on the test set. One can see that in the *cross-silo* settings, FedAvg will prioritize client $5, 7$ over other clients with a probability (weighting) of 0.18 and 0.17, respectively, resulting in sub-optimal performance. On the other hand, client 1 would only obtain a probability (weighting) of 0.11, thus down-weighting the participation of client 1 during FL pre-training.

Second, the introduction of *an angular measure of divergence* in FedAvg (i.e., L-DAWA$_{FedAvg}$) resulted in similar performance as with L-DAWA, for all the clients. One can see from Table 2, that clients aggregation with L-DAWA and L-DAWA$_{FedAvg}$ show similar results suggesting δ effectively reduces the divergence, the effects of biased-weighting of FedAvg, and controls the optimization trajectory. We conjecture that this is due to L-DAWA and L-DAWA$_{FedAvg}$ trajectories being closer to each other in the optimization surface. On average, one can see that the introduction of δ in FedAvg provides nearly 2.45% gain in the performance across the clients.

1.6. L-DAWA in linear fine-tuning

In this section, we provide additional results and detailed analysis for linear evaluation of our proposed methods, which are excluded from the main text of our manuscript due to the page limit.

1.6.1 Cross-silo performance

We compare L-DAWA with FedAvg [5] , Loss [2], FedU [8], and EUC [4] on CIFAR-10 under *cross-silo* (K=10) FL settings (Table 3). Note that, for the EUC method, we use the 'layer-wise unit model discrepancy' measure to make a decision about the update of the global model's layer with the client's model layer during aggregation. One can see from Table 3 that L-DAWA obtains the highest performance under different local epochs for both SimCLR and Barlow Twins.

We further provide additional results on Tiny ImageNet

	SimCLR			Barlow Twins		
Method	E=1	E=5	E=10	E=1	E=5	E=10
FedAvg (Baseline)	50.92	65.42	71.07	51.65	58.84	65.02
Loss [2]	50.99	63.83	71.34	48.24	54.64	57.12
FedU [8]	51.35	64.63	70.36	50.60	58.26	64.55
EUC [4]	51.23	64.10	70.51	51.16	58.76	63.60
L-DAWA	**60.29**	**70.65**	**75.60**	**54.84**	**65.07**	**69.31**

Table 3: Linear-probe accuracy on downstream for FedAvg, FedU, Loss, EUC, and L-DAWA. Each method is pre-trained with SimCLR/Barlow Twins on the Non-iid version (α=0.1) of CIFAR-10 for R=200 rounds under the *cross-silo (K=10)* settings.

	SimCLR			Barlow Twins		
Method	E=1	E=5	E=10	E=1	E=5	E=10
FedAvg (Baseline)	16.87	27.70	32.92	8.03	12.50	15.40
Loss [2]	15.25	28.47	33.37	8.26	7.97	12.24
FedU [8]	16.41	27.46	32.63	8.62	13.40	15.16
EUC [4]	15.99	28.52	32.15	1.11	0.5	0.5
L-DAWA	**23.12**	**31.97**	**37.72**	**10.76**	**17.34**	**21.47**
L-DAWA$_{FedAvg}$	22.39	32.65	36.97	9.23	17.20	21.58
L-DAWA$_{Loss}$	21.83	31.74	37.35	7.94	6.92	11.90
L-DAWA$_{FedU}$	23.40	31.98	37.12	11.30	17.07	21.80

Table 4: Linear-probe accuracy on downstream for FedAvg, FedU, Loss, EUC, and L-DAWA. Each method is pre-trained with SimCLR/Barlow Twins on the Non-iid version (α=0.1) of Tiny-ImageNet for R=200 rounds under the *cross-silo (K=10)* settings.

Method	SimCLR			Barlow Twins		
	100%	1%	10%	100%	1%	10%
FedAvg	32.92	12.66	24.54	15.40	5.21	11.35
Loss	33.37	12.77	24.12	12.24	3.14	6.72
FedU	32.63	12.11	24.23	15.16	5.29	10.49
L-DAWA	37.22	12.74	26.31	21.47	**8.20**	15.19
L-DAWA$_{FedAvg}$	36.97	12.77	26.32	21.58	8.05	**15.82**
L-DAWA$_{Loss}$	37.12	**13.48**	**26.63**	11.90	3.58	6.87
L-DAWA$_{FedU}$	**37.35**	12.77	26.09	**21.80**	7.70	14.78

Table 5: Linear-probe evaluation on Tiny ImageNet for our proposed methods compared to SOTA baselines. Each method is pre-trained with SimCLR/Barlow Twins on the Non-iid version (α=0.1) of Tiny ImageNet for 200 rounds under the *cross-silo (K=10)* settings.

by comparing L-DAWA, L-DAWA$_{FedAvg}$, L-DAWA$_{Loss}$, L-DAWA$_{FedU}$ against FedAvg, Loss, and FedU. We pre-train SimCLR and Barlow Twins on Tiny ImageNet in *cross-silo* FL settings with 10 clients for 200 rounds with 10 local epochs per round. Table 5 shows the results of fine-tuning when 100%, 1%, and 10% Tiny ImageNet training data are available. One can see that even in the case of limited data, L-DAWA, and its variants provide substantial performance improvement compared to FedAvg, Loss, and FedU.

1.6.2 Cross-device performance

Although our analysis is mainly limited to the *cross-silo* settings, we also provide results for the more challenging *cross-device* settings. Table 6 shows that our proposed methods still obtain the best performance in most of the setup, except for the 10% training data settings on CIFAR-100 dataset. Especially, L-DAWA achieves significant gains in the extreme semi-supervised settings with only 1% training data on CIFAR-100. Additionally, when layer-wise divergence is introduced in FedAvg, Loss, and FedU, we see a performance improvement in most of the cases, as shown in Table 6, suggesting the importance of the integration of divergence into aggregation within *cross-device* setting.

1.7. Transfer learning on CIFAR10/100

We further evaluate the generalization of the learned features from FL pre-training by fine-tuning the resulting model on a different dataset. Such evaluation helps in assessing whether the pre-trained features can be transferred to different downstream tasks. We follow the same procedure that is adopted for linear evaluation. Specifically, we first perform FL pre-training on CIFAR-10 (CIFAR-100) followed by linear-probe (fine-tuning the last classification layer) on CIFAR100 (CIFAR10). Note that the CIFAR-10 classes and CIFAR-100 classes are mutually exclusive [3].

One can see from Table 7, that L-DAWA generalizes well for both SimCLR and Barlow Twins compared to other aggregation strategies in the *cross-silo* and *cross-device* settings. We further show that when the layer-wise divergence is introduced in FedAvg, Loss, and FedU, we find a performance improvement for these methods in most cases in *cross-silo* and *cross-device* settings.

1.8. Evaluation on federated supervised training

Our proposed method has the potential to be extended to the setting of federated supervised training. We conduct an evaluation on the Non-iid version of CIFAR-10 under *cross-silo* (K=10) setting. One can see from Table 8 that L-DAWA surpasses all other baseline methods (FedAvg, FedYogi and FedProx) by at most 1.41%.

1.9. Ablation study

1.9.1 Effectiveness of δ in L-DAWA

We find that the *angular measure of divergence* (i.e., δ) between the local clients model and the global model plays an important role in determining the trajectory of the final global model. In Table 9, we show that without δ, L-DAWA results in sub-optimal performance. The results in Table 9 imply that for prolonged training in FL, both FedAvg and L-DAWA (without δ) will result in sub-optimal performance. We also find that without δ, L-DAWA reduces to FairAvg [6], (i.e., $w_g^{r+1} = \frac{1}{K}\sum_{k=1}^{K} w_k^r$). However, one can see

Method	CIFAR-10						CIFAR-100					
	SimCLR			Barlow Twins			SimCLR			Barlow Twins		
	100%	1%	10%	Linear	1%	10%	100%	1%	10%	100%	1%	10%
FedAvg	68.66	52.83	66.22	62.07	44.50	57.60	44.59	14.18	**32.55**	32.65	8.53	20.85
Loss	66.09	48.93	63.24	56.40	40.31	52.02	44.83	14.05	32.40	33.27	9.11	**21.88**
FedU	68.52	51.52	66.20	61.43	45.17	57.01	44.56	13.54	31.86	32.89	9.06	21.61
L-DAWA	68.20	51.45	64.71	58.25	41.86	53.26	45.04	**14.64**	32.07	**34.12**	**9.39**	21.84
L-DAWA$_{FedAvg}$	**69.92**	52.15	65.73	**62.32**	44.75	56.93	44.19	13.79	31.80	33.20	8.46	21.78
L-DAWA$_{Loss}$	68.79	**53.66**	65.68	61.36	44.77	56.88	**45.08**	14.63	31.67	31.93	8.40	20.25
L-DAWA$_{FedU}$	69.69	52.41	**66.63**	62.19	**46.64**	**58.46**	44.97	14.05	31.85	32.84	8.95	21.07

Table 6: Comparison of the proposed aggregation strategy with state-of-the-art methods on CIFAR-10 and CIFAR-100 under *cross-device (K=100)* settings.

	Cross-silo				Cross-device			
	CIFAR-10 → CIFAR-100		CIFAR-100 → CIFAR-10		CIFAR-10 → CIFAR-100		CIFAR-100 → CIFAR-10	
Method	SimCLR	Barlow Twins	SimCLR	Barlow Twins	SimCLR	Barlow Twins	SimCLR	Barlow Twins
FedAvg	44.28	34.34	66.82	60.28	44.94	37.77	67.48	**57.71**
Loss	40.72	28.25	67.02	58.66	41.33	32.06	66.91	56.61
FedU	40.49	33.43	67.66	59.32	44.28	37.21	67.12	56.53
L-DAWA	46.31	**40.00**	**74.52**	**66.39**	43.65	33.51	**68.21**	57.66
L-DAWA$_{FedAvg}$	**46.38**	39.93	74.30	65.63	**45.28**	36.88	67.01	57.45
L-DAWA$_{Loss}$	46.19	38.74	74.26	66.03	45.07	37.18	67.99	55.55
L-DAWA$_{FedU}$	46.30	38.78	73.89	65.89	45.10	**38.08**	67.69	56.73

Table 7: Transfer learning on CIFAR10/100 under *cross-silo* and *cross-device* settings.

Aggregation Type	E1	E5	E10
FedAvg	77.91	83.76	81.31
FedYogi	77.49	72.50	74.85
FedProx	80.55	74.87	72.24
L-DAWA	**81.96**	**84.68**	**82.35**

Table 8: Supervised evaluation on the Non-iid version of CIFAR-10 under the *cross-silo (K=10)* settings. The models are trained for 500, 100 and 50 FL rounds corresponding to the settings of 1, 5 and 10 local epoch(s), respectively.

Method	E1	E5	E10
FedAvg	50.92	65.05	71.07
L-DAWA w/o δ	50.33	64.09	70.31
L-DAWA w/ δ	**60.29**	**70.65**	**75.60**

Table 9: Ablation study of δ: Each method is pre-trained with SimCLR on the Non-iid version of CIFAR-10 under the *cross-silo (K=10)* settings for $R = 200$ rounds.

in Table 9 that even treating all the clients with the same weighting (i.e., $\frac{1}{K}$), the results are sub-optimal.

Agg.strategy	SSL-Method	Momentum	E1	E5	E10
FedAvg	F-SimCLR	✓	**51.29**	**68.25**	**74.50**
	F-Barlow Twins	✓	**57.49**	63.49	66.56
	F-SimCLR	✗	49.05	59.52	66.36
	F-Barlow Twins	✗	53.36	**65.89**	**68.03**
Centralized	SimCLR	✓		**85.27**	
	Barlow Twins	✓		**81.55**	

Table 10: Ablation study of momentum: Linear evaluation accuracy on the CIFAR10 dataset with *cross-silo* settings and iid data.

1.9.2 Effects of momentum

We provide an ablation study in Table 10 to highlight the importance of SGD momentum during FL pre-training. In short, we find that turning on the SGD momentum for Barlow Twins during FL pre-training can adversely affect the downstream task performance. In contrast, SimCLR improves the downstream task performance by turning on the SGD momentum during FL pre-training. We conjecture that this is due to the less divergence caused by SimCLR compared to Barlow Twins in FL settings.

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] Yan Gao, Titouan Parcollet, Salah Zaiem, Javier Fernandez-Marques, Pedro PB de Gusmao, Daniel J Beutel, and Nicholas D Lane. End-to-end speech recognition from federated acoustic models. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7227–7231. IEEE, 2022. 3, 4

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

[4] Sunwoo Lee, Tuo Zhang, Chaoyang He, and Salman Avestimehr. Layer-wise adaptive model aggregation for scalable federated learning. *arXiv preprint arXiv:2110.10302*, 2021. 3, 4

[5] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017. 3

[6] Umberto Michieli and Mete Ozay. Are all users treated fairly in federated learning systems? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2318–2322, 2021. 4

[7] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021. 1

[8] Weiming Zhuang, Xin Gan, Yonggang Wen, Shuai Zhang, and Shuai Yi. Collaborative unsupervised visual representation learning from decentralized data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4912–4921, 2021. 3, 4