Federated Self-supervised Learning for Video Understanding

Anonymous ECCV submission

Paper ID 7693

1 Supplementary Material

1.1 Loss Landscape Analysis of Pretext Tasks

In this section, we provide further details of the loss surface around the model pretrained with VCOP, Speed, and CtP video self-supervised learning (video-SSL) approaches, both in centralized and Federated Learning (FL) settings as shown in the Fig. 1. We use filter normalization [2] method to generate these loss surfaces around the pretrained model. One can see from Fig. 1, that the video-SSL models pretrained with FL provide a wider minima than the centralized approach. Except for CtP, the loss surfaces of VCOP and Speed are populated by the local peaks both in centralized and FL settings. The heights of these local peaks are directly proportional to the loss values. These local peaks further hint toward the better retrieval performance obtained with CtP as it shows a fairly smooth loss surface compared to those obtained by VCOP and Speed, both in centralized and FL settings. We can further see from Fig. 1 that the local peaks in FL version of VCOP are higher than the corresponding peaks in the centralized version, suggesting the reasons of obtaining low retrieval performance in FL settings against the centralized settings.

1.2 Key Hyper-parameters Tuning

In this section, we show the tuning process for key hyper-parameters and how they affect retrieval performance.

Retrieval performance of FedVSSL for different values of α. Table. 1 shows the results of various combination of loss-based aggregation and FedAvg for FedVSSL. One can see that the retrieval accuracy is highest when the weighting for loss-based aggregation (α) equals 0.9. We keep $\alpha = 0.9$ for all the experiments in our main paper.

Effect of momentum on the performance of FedVSSL. Table 2 shows the effect of momentum on retrieval accuracy. Indeed, the SSL models perform better when eliminating momentum during local training.

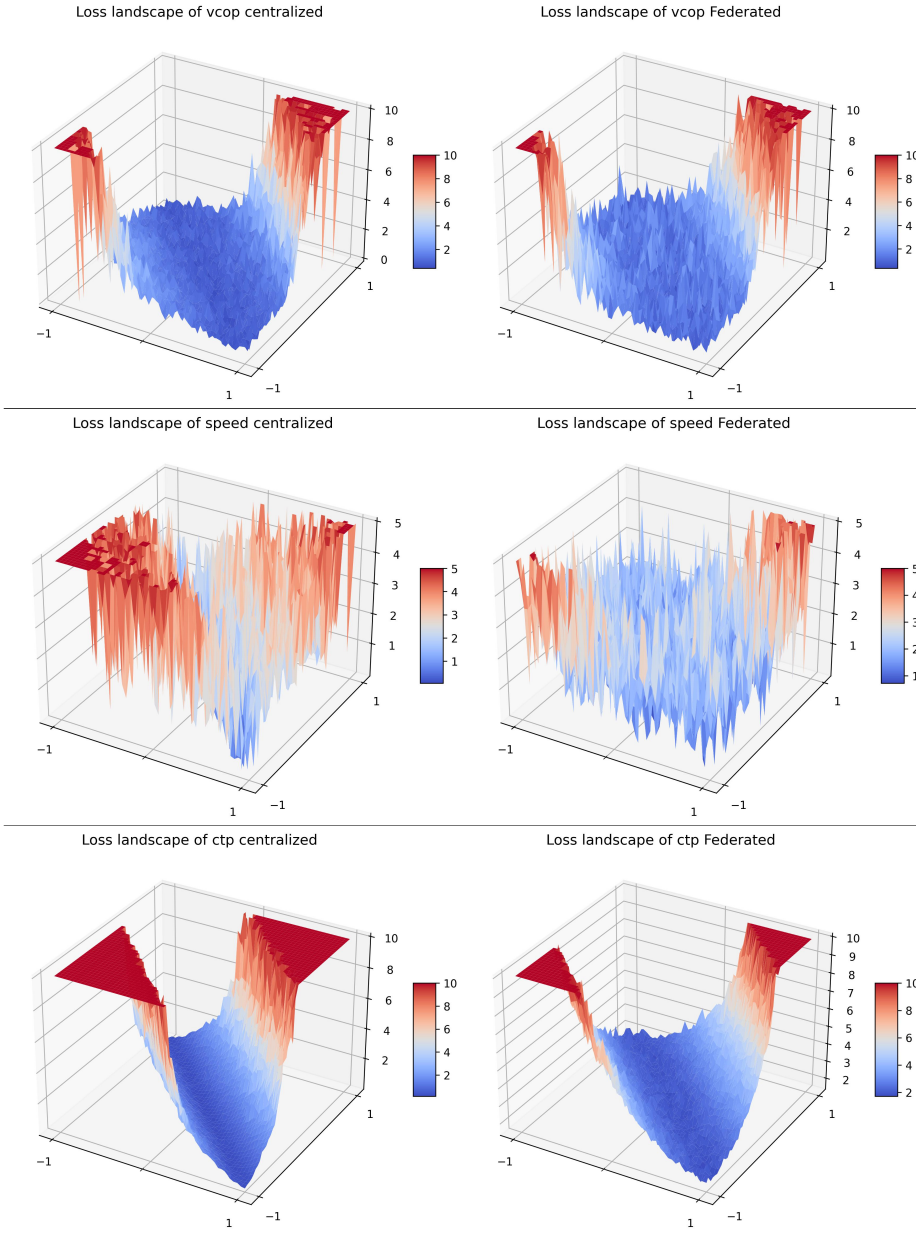

Fig. 1: Loss landscape around the model pretrained with VCOP, Speed and Ctp video-SSL approaches in centralized and FL with FedAvg. We can see that FL setting provide a wider loss surface than the centralized one.

Table 1: Retrieval accuracy of the combination of loss-based aggregation and FedAvg on UCF-101 and HMDB51 Dataset with 300 rounds pre-training on K400-Non-IID. α represents the weighting for loss-based aggregation.

α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
UCF101									
R@1	31.62	31.62	30.80	31.88	31.30	30.72	32.28	30.56	**32.65**
R@5	50.09	50.01	48.45	50.36	49.62	48.51	50.17	49.82	**50.49**
HMDB51									
R@1	14.84	15.75	16.01	15.69	14.97	16.21	**16.47**	15.56	16.41
R@5	33.92	36.54	34.58	34.71	33.86	34.71	35.36	35.16	**37.32**

Table 2: Retrieval accuracies (%) on UCF101 and HMDB51 with CtP[4] video-SSL approach using FedAvg, Loss-based aggregation and FedVSSL. The pre-training is performed on K400 (Non-IID). † means updating both θ^b and θ^{pt} in FedVSSL.

Method	Retrieval			
	UCF		HMDB	
	R@1	R@5	R@1	R@5
w/ momentum				
FedAvg(Baseline) [3]	29.29	48.90	13.66	32.42
Loss[1]	31.30	49.48	14.31	34.84
FedVSSL† $(\alpha = 0, \beta = 1)$	30.16	49.49	14.71	34.71
FedVSSL† $(\alpha = 1, \beta = 1)$	30.53	48.67	15.29	34.71
w/o momentum				
FedAvg	32.62	50.41	**16.54**	35.29
Loss	32.54	50.01	14.44	34.97
FedVSSL† $(\alpha = 0, \beta = 1)$	32.22	50.17	14.18	**36.80**
FedVSSL†$(\alpha = 1, \beta = 1)$	**32.67**	**50.28**	16.21	**36.80**

1.3 Qualitative Results of Video Clip Retrieval

In this section, we show the visual retrieval results of different FL aggregation strategies as well as the proposed FedVSSL approach. We show these results in Fig. 2 to Fig. 8. Fig. 2 shows the query clips in the test set of UCF-101, which are used to retrieve the Top-3 clips, Fig. 3 to Fig. 8, in the training set of UCF-101 using KNN. It can be seen from the visual results that our proposed FedVSSL approach performs comparatively better than FedAvg, Loss and FedU by retrieving similar clips in the category of BenchPress, CricketShot, and ThrowDiscus.

Fig. 2: Original query images from UCF-101 test split-1

Fig. 3: Top-3 retrieval results with FedAvg

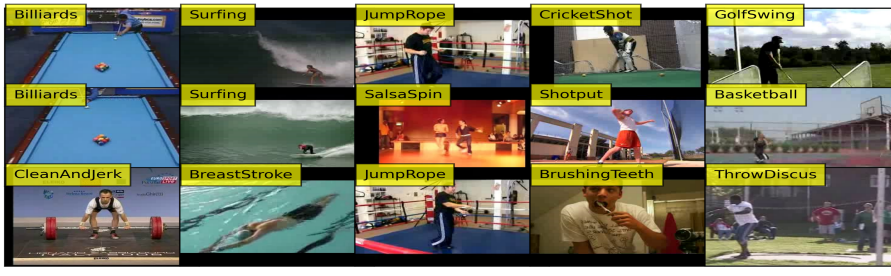

Fig. 4: Top-3 retrieval results with Loss

Fig. 5: Top-3 retrieval results with FedU

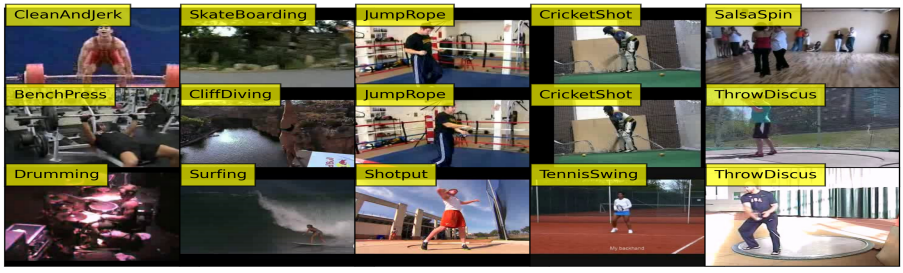

Fig. 6: Top-3 retrieval results with FedVSSL ($alpha = 0, \beta = 0$)

Fig. 7: Top-3 retrieval results with FedVSSL ($alpha = 0, \beta = 1$)

Fig. 8: Top-3 retrieval results with FedVSSL ($alpha = 1, \beta = 1$)

1.4 Further Exploration

Distributed Training of Video-SSL Approach on the Central Server with FedAvg. We first evaluate the performance of a centralized distributed model where the whole data is available, but the training is conducted in a

distributed fashion. In this settings, we first pretrain 5 models using CtP that share the data but not the weights. During each round, we randomly shuffle the whole dataset and divide it into 5 equal partitions. Each model is then trained on one of the partitions for 1 epoch, after which we performed the weighted averaging of all these models using FedAvg. Table 3 shows the result of this approach against the CtP with FedAvg and Non-IID data. We can see that training video-SSL in FL is equivalent to learning a large-scale distributed system with data privacy.

Table 3: Results of using FedAvg in the centralized distributed settings and FL settings

Method	Fine-Tune		Retrieval			
	UCF	HMDB	UCF		HMDB	
	Top-1	Top-1	R@1	R@5	R@1	R@5
FedAvg (Centralized)	81.76	49.85	30.16	48.27	14.71	32.68
FedAvg (FL)	81.95	49.15	29.29	48.90	13.66	32.42

Comparison between FedVSSL and FedU. In the FL settings with large number of clients, e.g., 100, FedU[5] is equivalent to only updating θ^b. This is because the clients are selected at random and the difference between the weights of the θ^b using FedAvg in the preceding round and current round will be quite large. This means that the classification head θ^{pt} on the client side is updated very few times with the global θ_g^{pt} throughout the training. Table 4 shows the fine-tuning and video clip retrieval results for CtP video-SSL approach using FedU and FedAvg (with only updating θ^b). One can see that FedU and FedAvg has obtained comparatively similar results indicating that the prediction network θ^{pt} on the clients are updated very few times.

Table 4: Comparison of FedU with FedAvg

Method	Fine-Tune		Retrieval			
	UCF	HMDB	UCF		HMDB	
	Top-1	Top-1	R@1	R@5	R@1	R@5
FedU	80.17	53.73	34.07	52.29	14.90	36.67
FedAvg(update θ^b only)	79.91	52.94	34.34	51.71	15.82	36.01

References

1. Gao, Y., Parcollet, T., Zaiem, S., Fernandez-Marques, J., de Gusmao, P.P., Beutel, D.J., Lane, N.D.: End-to-end speech recognition from federated acoustic models. arXiv preprint arXiv:2104.14297 (2021)
2. Li, H., Xu, Z., Taylor, G., Studer, C., Goldstein, T.: Visualizing the loss landscape of neural nets. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. pp. 6391–6401 (2018)
3. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. pp. 1273–1282. PMLR (2017)
4. Wang, G., Zhou, Y., Luo, C., Xie, W., Zeng, W., Xiong, Z.: Unsupervised visual representation learning by tracking patches in video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2563–2572 (2021)
5. Zhuang, W., Gan, X., Wen, Y., Zhang, S., Yi, S.: Collaborative unsupervised visual representation learning from decentralized data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4912–4921 (2021)