Supplementary Material: Ensemble Attention Distillation for Privacy-Preserving Federated Learning

Xuan Gong[1,2], Abhishek Sharma[2], Srikrishna Karanam[2], Ziyan Wu[2],
Terrence Chen[2], David Doermann[1], Arun Innanje[2]
[1]University at Buffalo, Buffalo NY [2]United Imaging Intelligence, Cambridge MA
{xuangong, doermann}@buffalo.edu {first.last}@uii-ai.com

1. Extension to Segmentation

Figure 1 shows the framework of FedAD when extended to segmentation. The class-specific attention maps are obtained through activation on the pixel-wise logits. The logits ensemble are conducted pixel-wisely. And the attention ensemble is to take class-specific intersection and union activation maps, and implemented in the same way as that in classification tasks.

2. Experiments

2.1. Implementation details for models trained on CIFAR-10/100

Following FedDF [1], the number of local nodes is set to $K = 20$. Random crop, flip, and cutout are used as data augmentation strategies. We train each local model individually with SGD and CosineAnnealing, decreasing the learning

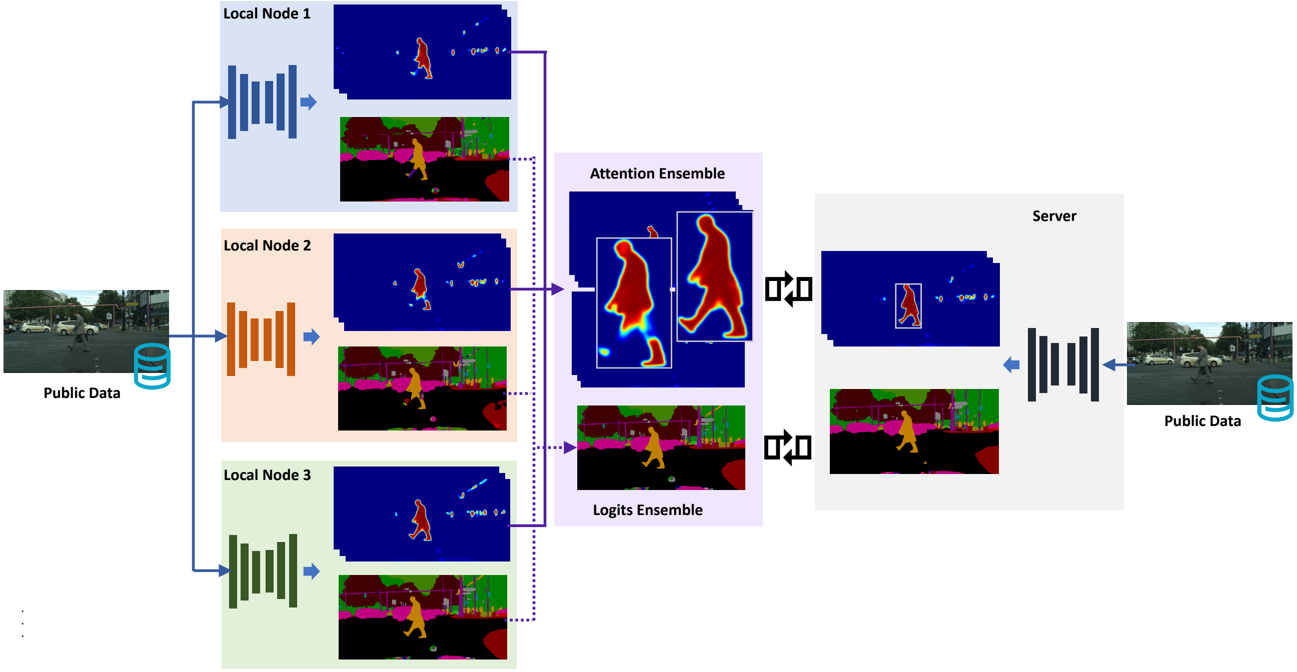

Figure 1. The framework of FedAD when extended to segmentation.

rate from 0.0025 to 0.001 in 500 epochs with a batch size of 16. For distillation, we use the Adam optimizer, a constant learning rate of 1e-3, and a batch size of 512. We train CIFAR-10 and CIFAR-100 with 200 and 10 epochs respectively. Weight decay is set to 3e-4 and 0 for local training and distillation respectively.

2.2. Implementation details for models trained on Chest X-Ray Images

For samples with multiple classes labeled as positive, we choose the most infrequent one (the class with least positive samples) as its label for the Dirichlet data split. We use ResNet-34 with batch size as 32 and the same data augmentation methods as in prior work [3]. Each local model is trained individually with SGD and CosineAnnealing and a decreasing learning rate from 1e-3 to 1e-6 across 20 epochs. For distillation, we use SGD and CosineAnnealing, and a decreasing learning rate from 1e-2 to 1e-3 across 20 epochs.

2.3. Extension to NLP tasks

We evaluate our framework on two text classification datasets: AG News [4] and SST2 [2] using the settings in FedDF [1]. The table below reports our FedAD's accuracy (%) on two text classification datasets: AG News and SST2 (using the same experimental settings as FedDF). We can note FedAD gives competitive performance on both datasets.

Dataset	FedAVG	FedDF	FedAD	Dataset	FedAVG	FedDF	FedAD
AG News	91.98	92.57	92.01	*SST2*	87.13	88.51	88.59

References

[1] Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. *34th Conference on Neural Information Processing Systems*, 2020. 1, 2

[2] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, 2013. 2

[3] Wenwu Ye, Jin Yao, Hui Xue, and Yi Li. Weakly supervised lesion localization with probabilistic-cam pooling. *arXiv preprint arXiv:2005.14480*, 2020. 2

[4] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing systems*, pages 649–657, 2015. 2