Supplementary Material for SysML 2020 Rebuttal

Program Chair's Approval

Dimitris Papailiopoulos 11/17/19 Hide DP
Re: [Systems and ML 2020] Author Feedback Period Extended to Nov 20
To: [redacted]
Cc: Systems and ML 2020 Program Chairs

Dear [redacted],

Yes you may use anonymous links and content as long as your identity is hidden. However, please note that the reviewers are not expected to use material from external links during their assessment.

Best,
Dimitris

Appendix 1: CIFAR10 with VGG11

We adopt a VGG11 architecture (Simonyan & Zisserman, 2014) and simulate the training of CIFAR10 (Krizhevsky et al., 2009) dataset using time measurements on Raspberry Pi 4's. The data are randomly partitioned to 5 clients. Every client updates the model with a mini-batch of 20 data, and 5 such mini-batches in a "federation". The samples are 1% (500) data randomly drawn from the training set. The sample data are excluded from the client data. We use an SGD (LR = 0.35) optimizer for this simulation. At each pruning "level", we remove 5% of parameters from the input and output layer, and 10% of parameters from the other layers. The model has 59.1%, 34.9%, 20.6%, and 12.2% parameters at levels 5, 10, 15, 20, respectively. In iterative pruning, we start from level 10, end in level 20, and prune once after every 200 federations, while in one-shot pruning, we keep it at level 20. The results are presented in Figure A.

Since VGG11 is primarily convolutional, and we are not able to find implementation of sparse convolutional kernels in any existing libraries, the time reduction mainly comes from communication, which is less evident than the reduction in LeNet-300-100 architecture. Nevertheless, we still find that Figure A(a) agrees with Figure 10, and Figure A(b) agrees with Figure 11 to a large extent. The reason why sample-less pruning does not work in Figure A(b) is possibly due to its damage to the network architecture (see Sec 6.1 for details) for the current pruning level. Sample-less pruning can possibly work if we prune to a lower level.

Appendix 2: FEMNIST with Conv-FEMNIST and Sampled Clients

The experimental setting is the same as in Section 5.2 (see Table 1), except that, instead of sampling all clients, we train with the data randomly sampled from 10 out of 193 clients in each federation. The sampled clients in different federations are generally different. We also updated the learning rate to 0.031 which was determined empirically. The results are presented in Figure B.

Because we sample from 10 out of 193 clients, the learning curve is slowed down. Nevertheless, our paper's conclusion still holds. Same as before, the reason why sample-less pruning does not work in Figure B(b) is possibly due to its damage to the network architecture (see Sec 6.1 for details) for the current pruning level; it can possibly work if we prune to a lower level.

Appendix 3: FEMNIST with LeNet-300-100

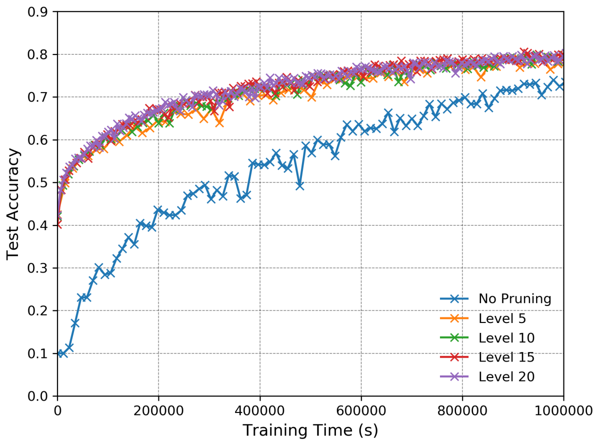
The experimental setting is the same as in Section 5.2, except that we use a much smaller LeNet-300-100 (see Table 1) for training. We set the pruning rates to be 0.1, 0.1, 0.05 respectively in each layer, and thus at levels 5, 10, 15, 20, there are 59.5%, 35.6%, 21.3%, 12.9% of the parameters left, respectively. The optimizer is SGD, LR = 0.05. In iterative pruning, we start from level 10, end at level 20, and prune once after 200 federations, while in one-shot pruning, we keep it at level 20. The results are presented in Figure C.

We consider the LeNet-300-100 architecture underparameterized for the FEMNIST dataset. Figure C shows that even with such an underparamterized model, our main claims in the paper still hold. The only difference is the lower accuracy at convergence (of the non-pruned model) compared with larger models such as Conv-FEMNIST that we used in Section 5.2.

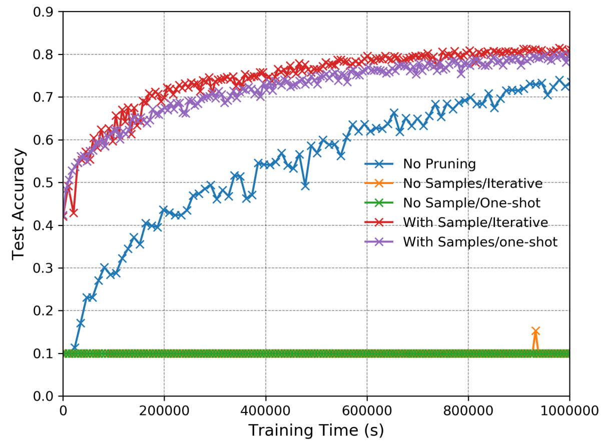
References

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

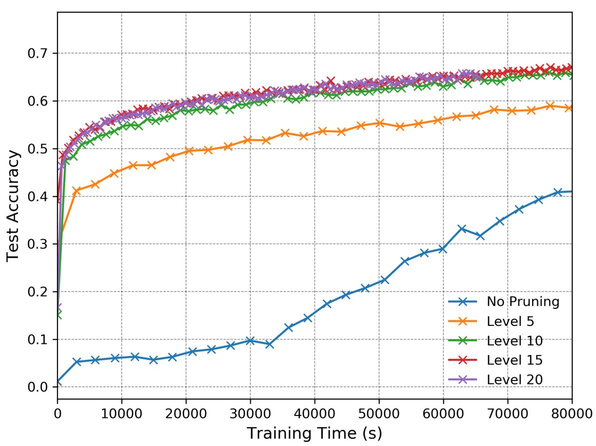


(a) Comparing Sample-based, One-shot Approach at Pruning Level 5, 10, 15, 20

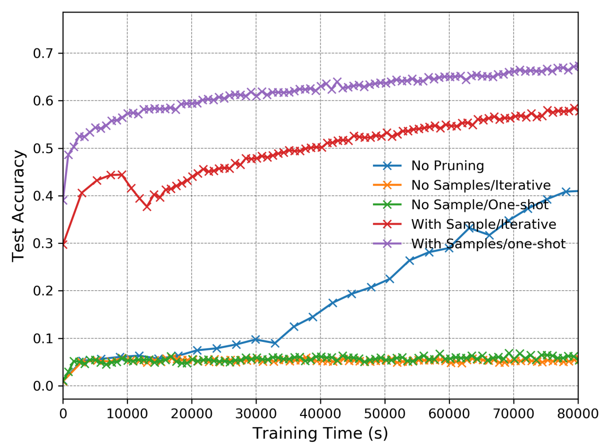


(b) The 4 Possible Pruning Cases

Figure A. CIFAR10 Dataset (i.i.d Partitioned) with VGG11 Architecture

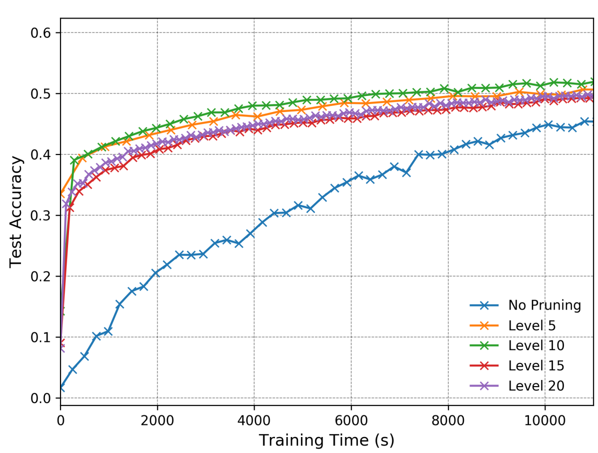


(a) Comparing Sample-based, One-shot Approach at Pruning Level 5, 10, 15, 20

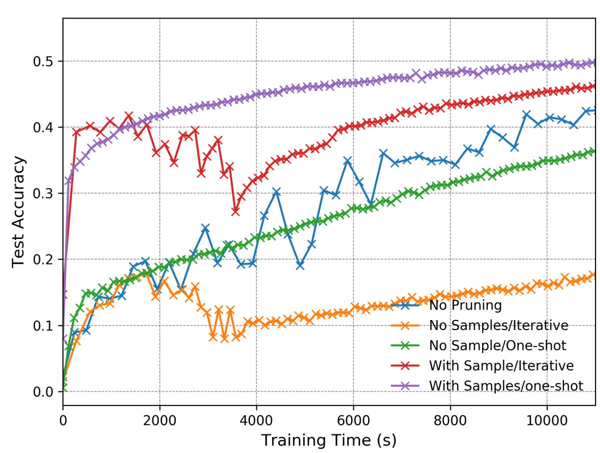


(b) The 4 Possible Pruning Cases

Figure B. FEMNIST Dataset (Non-i.i.d Partitioned) with Conv-FEMNIST Architecture and Sampled Clients



(a) Comparing Sample-based, One-shot Approach at Pruning Level 5, 10, 15, 20



(b) The 4 Possible Pruning Cases

Figure C. FEMNIST Dataset (Non-i.i.d Partitioned) with LeNet-300-100 Architecture