Pruning in federated learning

Hoang Trung Hieu

2020

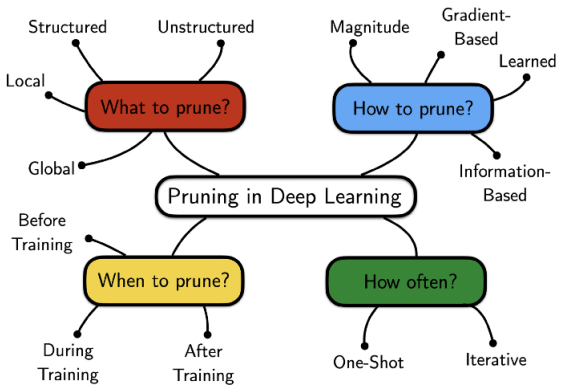
Abstract

1 Neural network pruning

To reduce the size of network there a quite a few techniques were introduced by removing parameters. In [1], they used the second-order Taylor expansion, but computing in Hessian matrix is unattainable to modern deep neural network. It shows that a fully trained dense network can be pruned to little parameters without degrading performance too much. In case of training a sparse sub-network, the lottery ticket hypothesis was introduced in [3], [4]. They believe that dense networks contain sparse sub-networks that can be trained to perform as good as the original dense model.
In gereral, in order to develop any competitive pruning technique, it requires to answer the following 4 questions:

- **What connectivity structures to prune?** *Unstructured pruning* does not consider any relationships between the pruned weights. *Structured pruning*, on the other hand, prunes weights in groups. *Local pruning* enforces that one prunes p percent of weights from each layer. *Global pruning*, on the other hand, is unrestricted and simply requires that the total number of weights across the entire network is pruned by p percent.
- **What is the pruning criterion?** A popular technique is magnitude-based pruning ([2]) that keep the large magnitude weights since it has more impact on the function fit and should be pruned less. In addition to this, there are technique which use gradient-based methods or even higher-order curvature information.
- **When we prune?** There are 3 stamps when can apply pruning: before (initialisation-based) , during (training-based) and after training. *Pruning before training* performs the opting based on the untrained weights. *Pruning during training*, on the other hand, is often associated with regularization and ideas of dropout. When *pruning after the training* has converged the performance often decreases, which makes it necessary to retrain/fine-tune and to give the network a chance to readjust.

- **How often to perform the pruning step?** *Iterative* procedures prune only a small number of weights after one training run but reiterate the train - score - prune - rewind cycle. On the other hand, *one-shot pruning* ([6]) performs only a single time at the end of training.



Figure[1] What, When, How and How often to prune?

In [3], they propose iterative magnitude pruning, in particular, unstructed, magnitude-based, iterative and initialisation-based prunning. In[5], they suggested the mixed-method which combines two factor magnitude and gradient sensitivity into one criterion by taking the multiplication.

2 Federated Learning pruning

References

[1] Yann LeCun, John S Denker, and Sara A Solla. *Optimal brain damage*. In Advances in neural information processing systems, pages 598–605, 1990.

[2] Song Han, Jeff Pool, John Tran, and William Dally. *Learning both weights and connections for efficient neural network*. In Advances in neural information processing systems, pages 1135–1143, 2015.

[3] Jonathan Frankle and Michael Carbin. *The lottery ticket hypothesis: Finding sparse, trainable neural networks*. In ICLR, 2019

[4] Ari Morcos, Haonan Yu, Michela Paganini, and Yuandong Tian. *One ticket to win them all: generalizing lottery ticket initializations across datasets and optimizers*. In Advances in Neural Information Processing Systems, pages 4933–4943, 2019.

[5] Dániel Lévai, Zsolt Zombori. *Data-dependent Pruning to find the Winning Lottery Ticket*, https://arxiv.org/abs/2006.14350, 2020

[6] Namhoon Lee, Thalaiyasingam Ajanthan, and Philip Torr. Snip: *Single-shot network pruning based on connection sensitivity.* In International Conference on Learning Representations, 2019. https://openreview.net/forum?id=B1VZqjAcYX .