MMI 712 Assignment 3

Network Pruning

1 Questions (40 pts)

Network pruning is a technique used to reduce the complexity of a neural network by removing unnecessary connections and parameters. The goal is to reduce the number of parameters in the model while maintaining or improving its time and space complexity.

1. Briefly explain "Lottery Ticket Hypothesis"[1] (∼500 words), highlighting the key concepts.
2. What are the trade-offs between pruning a network during training versus after training?
3. What are the effects of pruning other than reducing space and time complexity of a network?

2 Code (60 pts)

Find the notebook file named "pruning.ipynb" in the assignment folder. It includes the code for preparing CIFAR100 dataloader, a pretrained model along with the code for calculating top-5 accuracy. Implement 3 data-free pruning techniques $l_1, l_2, random$ and plot the pruning ratio vs. accuracy graph for 10 pruning ratios (0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9) for each pruning technique. Comment on the results.

3 Submission

Submission must include two files;

1. Assignment report in pdf format
2. Your final notebook with cell results on. Please remove the data before submission.

References

[1] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.