Benjamin Ebanks 18.065

Yajvan Ravan 04/02/24

**Neural Network Pruning: A Linear Algebra Analysis**

In recent years, the field of deep learning has gained a lot of traction, and as a result, neural networks have significantly increased in complexity, featuring more layers and parameters. While there are many benefits in having these complex networks, such as improved accuracy and the ability to model more intricate patterns in data, they also carry redundant parameters and sometimes lack interpretability. A popular fix to this issue is neural network pruning, a method that involves removing unnecessary weights without significantly decreasing the accuracy of the model. For our final project, we will be exploring three approximation techniques that we have learned thus far in class as methods of neural network pruning, including low rank approximation, principal component analysis, and randomized numerical linear algebra.

We want to use PCA to optimize the architecture design of these networks similar to [2]. PCA will allow us to prune layers by determining the optimal output dimensionality of each feature layer, similar to the method in [2], and then retraining. Another way we wish to use PCA is similar to [4], to reproject the weights of each layer & prune without retraining. The second method we want to explore is that of using low-rank approximations of weights matrices. Similar to [1] and [3], we wish to exploit low-rank approximations of layer weights to speed up multiplication & therefore overall performance. Finally, a third method we wish to explore is that of using approximate, or randomized, matrix multiplication to perform the same.

For each of these methods, we plan to implement them on well-known pre-trained neural network backbones, including AlexNet, ResNet, and DenseNet, & compare the performance on the MNIST dataset. We also wish to perform a theoretical analysis of performance improvement and compare this with empirical results. If time permits, we also want to measure performance across a range of domains beyond image classification.

The goal of this project is to compare and contrast various methods for pruning neural networks with dimensionality reduction methods. We expect that for this simple task, dimensionality reduction should perform very well, with PCA performing the best, followed by low-rank approximations. We expect that the final inference time of the pruned networks should be substantially smaller than original, with similar accuracy retained.

References

1. Emily Denton, Wojciech Zaremba, Joan Bruna, Yann LeCun, and Rob Fergus. 2014. Exploiting linear structure within convolutional networks for efficient evaluation. In Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'14). MIT Press, Cambridge, MA, USA, 1269–1277.
2. Garg, I., Panda, P. and Roy, K. (2020) ‘A low effort approach to structured CNN design using PCA’, IEEE Access, 8, pp. 1347–1360. doi:10.1109/access.2019.2961960.
3. Jaderberg, M., Vedaldi, A. and Zisserman, A. (2014) ‘Speeding up convolutional neural networks with low rank expansions’, Proceedings of the British Machine Vision Conference 2014 [Preprint]. doi:10.5244/c.28.88.
4. Levin, Asriel, Todd Leen, and John Moody. "Fast pruning using principal components." Advances in neural information processing systems 6 (1993).