Optimizing Neural Networks for Mobile Devices

Erdi Çallı

Radboud University Nijmegen, Neurant

Abstract

Studies show that pruning unnecessary neurons in an ANN is effective in reducing the model complexity. Activation based pruning, sums up the neuron activation counts. Using that information, it determines the useless neurons, and deletes them. In theory, this method reduces the training time per cycle, and prioritizes important connections. We set up an experiment to see the efficiency of this method. We defined a problem with a simple optimal solution, summation of two inputs. We defined a network with redundant parameters and trained it in different configurations. Our experiments show that this method is working to its expectations.

We can talk about how it is important in the nature, check Rust12, George, and Bolouri23 (1997)

Optimizing Neural Networks for Mobile Devices

Introduction

ANN's have several parameters such as number of hidden layers, number of neurons in a layer, or the structure of a layer. Until now, we have seen different combinations for these parameters. For example, Simonyan and Zisserman (2014) introduces a model called VGGNet. VGGNet introduces more layers (16 to 19) than the previous models. They show how this parameter effects the accuracy. He, Zhang, Ren, and Sun (2015) introduces the residual connections. This new connection between layers is capable of stacking more layers than before. Training up to 152 layers, they show superior accuracy. Zagoruyko and Komodakis (2016) compares having higher number of neurons in each layer to having more layers. Each combination resulting in a unique model with a different accuracy level. In contrast to all these, Szegedy et al. (2014) suggests something different. Having a good harmony within the network works better than having more parameters. Supporting that, Canziani, Paszke, and Culurciello (2016) does a detailed comparison of different models. Their findings show that increasing the number of hidden layers or the number of neurons in a layer does not necessarily increase the accuracy.

Following these,

References

Canziani, A., Paszke, A., & Culurciello, E. (2016, 05). An analysis of deep neural network models for practical applications. Retrieved from https://arxiv.org/abs/1605.07678

He, K., Zhang, X., Ren, S., & Sun, J. (2015, 12). Deep residual learning for image recognition. Retrieved from https://arxiv.org/abs/1512.03385

Rust12, A. G., George, S., & Bolouri23, H. (1997). Activity-based pruning in developmental artificial neural networks. In *Fourth european conference on artificial life* (p. 224).

Simonyan, K., & Zisserman, A. (2014, 09). Very deep convolutional networks for large-scale image recognition. Retrieved from https://arxiv.org/abs/1409.1556

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2014, 09). Going deeper with convolutions. Retrieved from https://arxiv.org/abs/1409.4842

Zagoruyko, S., & Komodakis, N. (2016, 05). Wide residual networks. Retrieved from https://arxiv.org/abs/1605.07146