CRUNCH Seminars at Brown, Division of Applied Mathematics

Friday - February 26, 2021

Sparsity in Deep Learning: Pruning and growth for efficient inference and training in neural networks

Zongren, Zou CRUNCH

The growing energy and performance costs of deep learning have driven the community to reduce the size of neural networks by selectively pruning components. Similarly to their biological counterparts, sparse networks generalize just as well, if not better than, the original dense networks. Sparsity can reduce the memory footprint of regular networks to fit mobile devices, as well as shorten training time for ever growing networks. In this paper, we survey prior work on sparsity in deep learning and provide an extensive tutorial of sparsification for both inference and training. We describe approaches to remove and add elements of neural networks, different training strategies to achieve model sparsity, and mechanisms to exploit sparsity in practice. Our work distills ideas from more than 300 research papers and provides guidance to practitioners who wish to utilize sparsity today, as well as to researchers whose goal is to push the frontier forward. We include the necessary background on mathematical methods in sparsification, describe phenomena such as early structure adaptation, the intricate relations between sparsity and the training process, and show techniques for achieving acceleration on real hardware. We also define a metric of pruned parameter efficiency that could serve as a baseline for comparison of different sparse networks. We close by speculating on how sparsity can improve future workloads and outline major open problems in the field.