CMP SCI 635: Modern Computer Architecture

SCNN: An Accelerator for Compressed-sparse Convolutional Neural Networks

Name: Kunjal Panchal Date: 3rd Dec, 2019 Student ID: 32126469

Paper: 22 – Data Parallel [Parashar17]

Strengths:

1. Dynamic pruning can produce a sparse matrix whose irregular data access patterns can negate the pruning benefits; thus, accelerators specialized for sparse matrices are major performance benefit.

2. The advantage is that if the actual convolution can still be computed efficiently because sparsity is introduced only at the outer loop of the convolution operation and we can still take the advantage of the continuous memory layout.

Weaknesses:

1. The topic of compressed NN is at the crossroads of many different applications with different goals and constraints. I feel like it's one of those subjects where research and implementations are completely disconnected. There are a huge number of techniques which reduce computation costs by orders of magnitude, but very few of them translate into real world speedup. Finding a way to bridge that gap could lead to widespread, immediate benefits.

2. Most of the papers on network compression I've came across consist of very different techniques, many with little to no theory to back up. I've also seen a few papers/reports showing that many compression techniques can be outperformed/matched by just training better/longer a smaller network, and so on. While there are a few prominent approaches (quantization, binarization, pruning, etc), it is not clear which one is most promising.

Questions/Assertions:

1. Is there a standard comparison metric? Compression techniques typically affect the inference time of the network, and some focus on decreasing bits/param instead of number of parameters, so it's not clear how these should be compared.

2. For the few compression papers I have seen, they have been applied to VGG16 and similar networks, which are not particularly parameter efficient compared to current state of the art. Generally, there have been a lot of progress in more efficient and compact CNN architectures the last years, for example MobileNets etc. A question is whether these new architectures are still significantly compressible, or if the gains have been eaten up?

3. If the goal is to reduce the model size, most of the existing compression methods (e.g. DeepCompression) make the network sparse by pruning some weights (e.g. removing the smallest 10% weights). Though this reduces the number of parameters, it doesn't translate to a smaller model. Even a zero valued weight consumes space.

4. From theoretical point of view, I haven't seen any study on understanding what really happened in pruning. I wonder what are this redundant information that got removed from the network? Is there a way to understand what really happened inside the network? Can we apply feature visualization technique and see what really changed? I feel this might give more information on networks.

5. FPGAs are better suited than GPUs for implementing inference in many DNN models if the following are true:
   - We're concerned about power consumption (the biggest advantage of using an FPGA)
   - The model parameters (after compression) are small enough to fit in on-chip memory (~5-30 MB)