Project 3: Quantisation of networks with learnt hashed nets

Problem statement

DNNs are over-parameterized and do not require high precision to store weights. This has led to different approaches to quantise the weights of DNNs. One approach maps weights in the network on to a select set of allowed weights. For instance, if we choose 16 allowed weights then we can hash every weight in the network to the closest allowed value. Then, each of the weights could be stored with only 4 bits referring to the allowed weight, while each weight can be a floating point number with high precision. Several approaches to do this involve iterative optimisation where a network is trained without constraints on the weights, then the weights are clustered, and each weight is mapped on to the nearest cluster center. Thus these cluster centers become allowed weights. This process decouples the training from the hashing.

This project explores the idea of hashing as part of the training process. This requires two changes: (a) adding additional variables for the allowed weights, and (b) modelling the hash function as a trainable layer in the network. Then the modified network can be trained with standard back-propagation and the allowed weights are learnt. For inference, the model is simplified to remove the additional layers and quantise each of the weights to the nearest allowed weight. The hypothesis is that such training based hashing would be more accurate / efficient than existing methods.

References

Hashed Nets -> [1504.04788] Compressing Neural Networks with the Hashing Trick
Deep-compression -> [1510.00149] Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding

Milestone 1

1. Propose a way to model hashing using existing operators in most frameworks such as deconvolution, max-pooling
2. Implement the hashing layer and check correctness for forward and backward propagation
3. Implement the training of a simple model such as Resnet34 for CIFAR 10 with hashed nets and evaluate accuracy

Milestone 2

1. Perform detailed comparison with results in existing literature (same models and datasets) both on accuracy and model size
2. Perform experiments to evaluate the right number of allowed weights to use and the impact on accuracy