MA 598 Machine Learning Seminar
Summary 1

ID 5107

January 18, 2019

1 A Survey of Model Compression and Acceleration for Deep Neural Networks

The paper surveys four methods that have been successfully used to reduce the complexity and size of deep neural networks (DNN for short). The four methods discussed are: parameter pruning and sharing, low-rank factorization, transferred/compact convolutional filters, and knowledge distillation.

1.1 Parameter pruning

Parameter pruning is the simplest method discussed in the paper. The idea behind it is to reduce the complexity of a DNN by removing parameters which are not crucial to the model. In the paper, three methods of parameter pruning are discussed: quantization and binarization, (more largely) pruning and sharing, and a structural matrix approach.

Quantization and binarization compresses the original DNN by reducing the number of bits used to represent the weights. This is achieved by first pruning unimportant connections in the DNN (retaining the sparsely connected ones) and then quantizing the link weights by using weight sharing. In the referenced paper, some further methods are applied, but we omit them here for briefness.

The extremal case in this direction is called binarization, where we have a one bit representation of each weight. This is exactly what is done in, e.g, BinaryConnect, BinaryNet, etc. The drawback to this approach is that the

accuracy of binary networks is lowered. Moreover, binarization schemes are based on simple matrix approximations and ignore the effects of binarization on accuracy.

1.2 Low-rank factorization

The second method discussed involves reducing the number of convolution operations done in deep convolutional neural networks (CNN for short). This can be achieved in a variety of ways, but most prominently by applying low-rank filters. In the paper, two such filters (or decomposition methods) stand-out–the canonical polyadic decomposition (CP for short) and the batch normalization (BN for short). It is mentioned that finding the best low-rank approximation in the CP decomposition is an ill-posed problem which may not have a solution, but for BN a solution always exists. One of the drawbacks of this approach lies in its computational cost since it involves decomposition operations.

1.3 Transferred/compact convolutional filters

The idea behind using transferred convolution filters is motivated by recent incorporation of equivariant group theory into the study of CNN. It is mentioned in the paper that this approach is currently lacking in theory, but there is a strong empirical evidence to support the notion that CNN posses a translation invariance of the following sort: If T is a transform matrix, x an input, and Φ a network,

$$T\Phi(x) = \Phi(Tx). \tag{1}$$

1.4 Knowledge distillation