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 presented surveys four methods of compressing the size and complexity of DNN. The four methods discussed are: parameter pruning, low-rank factorization, transferred/compact convolutional filters, and knowledge distillation. It is important, for practical applications, to have a small network, because of physical constraints on many devices including distributed systems, embedded devices, and FPGA.

Parameter perhaps the most conceptually simple method discussed in the paper. The idea behind it is this: to achieve a more compact DNN remove the non-crucial parameters from the model. Three methods of achieving this type of compression are discussed: quantization and binarization, pruning and sharing, and a structural matrix approach.

Of particular interest is binarization, where we force a one-bit representation of each weight as is done in successfully in several DNN such as BinaryConnect, BinaryNet, and XNORNet. This achieves a high degree of compression, but at a cost in accuracy on very large CNN. However, progress is being made in this direction, as there is strong empirical evidence that networks trained with back-propagation are resilient to specific weight distortions, in particular, binary weight.

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.

The third method discussed, 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,

$$T\Phi(x) = \Phi(Tx); \tag{1.1}$$

i.e., transforming the input x by T is the same as passing it through the network (or layer) Φ and then transforming the output by T. The idea is to apply certain transforms T to a small set of base filters. It is mentioned that this method achieves reduction in parameters with little drop in classification accuracy.

The last method, knowledge distillation, follows a student-teacher approach where the teacher network is trained from scratch and the student network is penalized according to softened versions of the teacher's output. One of the drawbacks of this method can only be applied to classification tasks with a soft-max loss function.