Distilling the Knowledge in a Neural Network

First week from machine learning workshop series

Hosein Fooladi

September 2, 2019

Contents

1 Main topic

Model compression in one of the hot topic today in machine learning community. One of the main purpose of model compression is being able to run model and make inference on edge device like mobile phones.

In this week, we focus on reading a paper with the title "Distilling the Knowledge in a Neural Network". The model should be small enough to fit on the memory of edge device, It should have low latency (the inference should be fast, which means the number of computations should be relatively small). For satisfying these constraints, we can not deploy the large model we trained on huge computational power on edge device. We should find a way compress the model, and at the same time, keep the accuracy high. In the paper it has been stated that:

"Many insects have a larval form that is optimized for extracting energy and nutrients from the environment and a completely different adult form that is optimized for the very different requirements of traveling and reproduction."

This emphasizes on the different requirements and objectives during training and test phase.

Generally, I can say there are three active area of reasaerch in order to compress complex and huge models and make them suitable for running on edge device:

- Weight pruning
- Quantization
- Knowledge distillation

In this session, we talked more about knowledge distillation and read the "Distilling the Knowledge in a Neural Network" more carefully. They introduces some tricks to make this technique works (e.g., introducing softmax temperature and dark knowledge). This paper can be considered as the main reference for knowledge distillation.

2 Side topics

During the session, we talked briefly about some other papers and topics that are relevant to the main paper. So, I am mentioning these topics here and anybody who is interested can learn more about them.

- Improved Knowledge Distillation via Teacher Assistant: Bridging the Gap Between Student and Teacher
  This paper claims that if the gap between capacity of large and small network becomes huge, the student network can not recover the accuracy of teacher networks. So, It introduces a concept "teacher assistant" to alleviate this issue.

- Introducing DistilBERT, a distilled version of BERT. Distill Bert is an attempt to use knowledge distillation to compress the BERT model and at the same time achieve relatively same accuracy. You can read this blog post to learn more.

- Quantized Neural Networks: Training Neural Networks with Low Precision Weights and Activations
  It is possible to quantize both weight and activation of the neural network, and at the same time achieve relatively high accuracy. It is even possible to use 1-bit quntization and surprisingly, have a good performance. Quantization can have a regularization effect and can make the model generalize better. Read this paper to learn more.

- Dark knowledge. Good paper and video is available.

3 Future reading

In this section, I am going to provide some useful code base, frameworks or materials for further reading to gain comprehensive understanding.

- Distiller: Distiller is an open-source Python package for neural network compression research. It is a very good package written with PyTorch that help you implement and experiment with distillation easily. you can learn more by reading the documentation.

- Implementation of "Improved Knowledge Distillation via Teacher Assistant: Bridging the Gap Between Student and Teacher" paper. The implementation and codes are well documented and you can easily reproduce the paper results.

- Tensorflow quantization guide for mobile & IoT. This is a very good resource to learn how you can apply quantization methods on the saved model and deploy it on the edge device.

4 Miscellaneous

I want these workshop series be useful for the audience, and in addition, I learn from you during presentation. So, please let me know your thought about organization, topics, or everything you think it is good to be considered.