Notes

Net2Net: ACCELERATING LEARNING VIA KNOWLEDGE TRANSFER

Authors of the Paper: Tianqi Chen, Ian Goodfellow, Jonathon Shlens

Compiled By: *Mohit Jain*

Discussion held on 1st August 2018

Summary

This paper introduces a new way of pre-training by introducing a *"teacher"* and a *"student"* network. The idea is to use a **function preserving transformation** to initialize the student network by using the weights of the teacher network. This ensures that the student network performs just as well as the old teacher network from the start and further training will just improve its performance.

The *Net2Net* procedure can used to increase the width and/or depth of the student network compared to the teacher network.

Net2WiderNet

This allows a layer to be replaced with a wider layer, meaning a layer that has more units. For convolution architectures, this means the layers will have more convolution channels.

Net2DeeperNet

This allows us to transform an existing net into a deeper one. This is done by inserting an identity transformation after any layer into the existing original network. This method is a special case of factoring a layer in the network. *Net2DeeperNet* essentially factorises an existing layer $L^{(i)}$ to $\boldsymbol{I}$ and $L^{(i)}$, where $\boldsymbol{I}$ is the identity mapping layer.

Effectiveness of Net2Net

Due to the function preserving method used, the new larger network will perform at least as well as the original smaller network. Also, networks trained using *Net2Net* converge faster to the same accuracy as networks of the same size initialised randomly.

One crucial point to note is that the final accuracy is independent of the training procedure used and only depends on the network. Hence, *Net2Net* doesn't help boost up the accuracy but only *quickens* the time taken to reach it!

The same hyperparameters used to train the smaller teacher network from scratch can also be used to train the larger student network using *Net2Net*. This makes the whole process simpler and less tedious. However, it is advised to set the initial learning rate of the student network to be 0.1 times the initial learning rate of the teacher network as the training of the new network can be considered as a continuation from the teacher network's training.

Shortcomings of Net2Net

1. *Net2Net* can only be used to increase the width and depth of the network. Therefore, kernel sizes in CNNs cannot be made using this method.
2. While increasing the depth of the network, *Net2Net* can only factorise a layer using identity mapping. A general non-identity mapping factorisation cannot be used.
3. *Net2Net* only works for idempotent activation functions; *an idempotent function,* Φ, *satisfies* $\Phi \circ \Phi = \Phi$, *such as ReLU.*

Link to the paper: *https://arxiv.org/abs/1511.05641*

You can also refer to this **blog** for more: https://mohitjain.me/2018/07/19/net2net/