周学习总结

许典

李宏毅视频

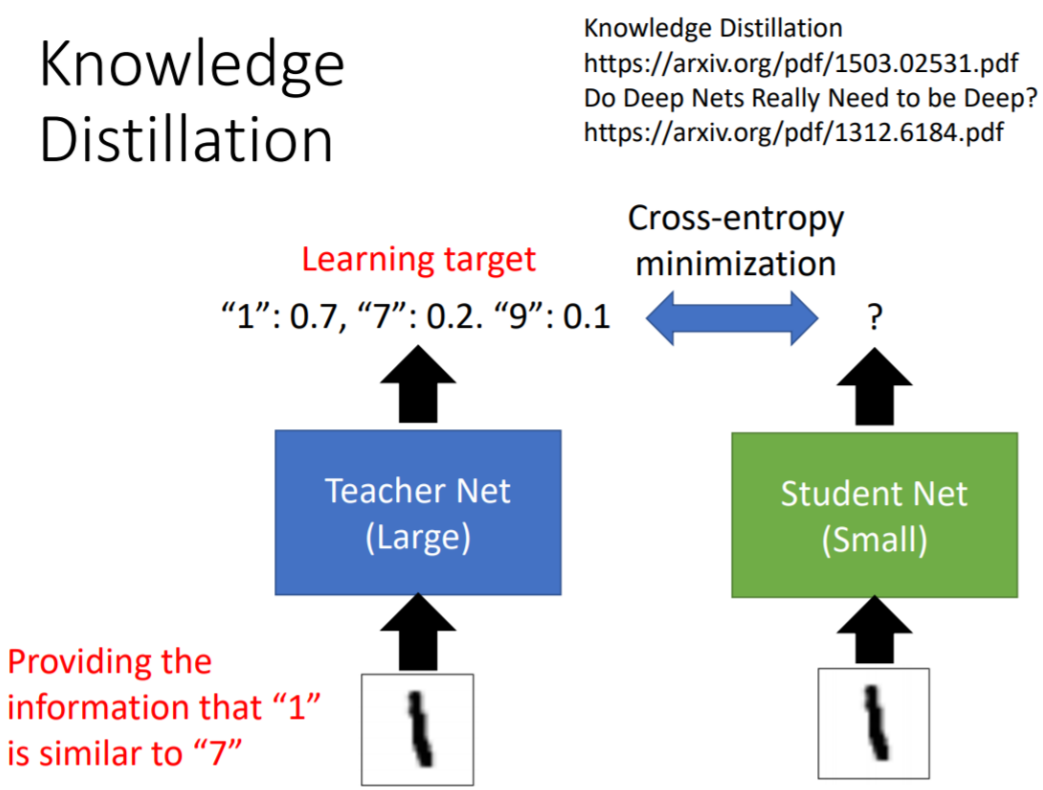
第七节：Network Compression

里面提到了四种网络压缩方法：

- 知识蒸馏 Knowledge Distillation
- 网络剪枝 Network Pruning
- 用少量参数来做 CNN Architecture Design
- 参数量化 Weight Quantization

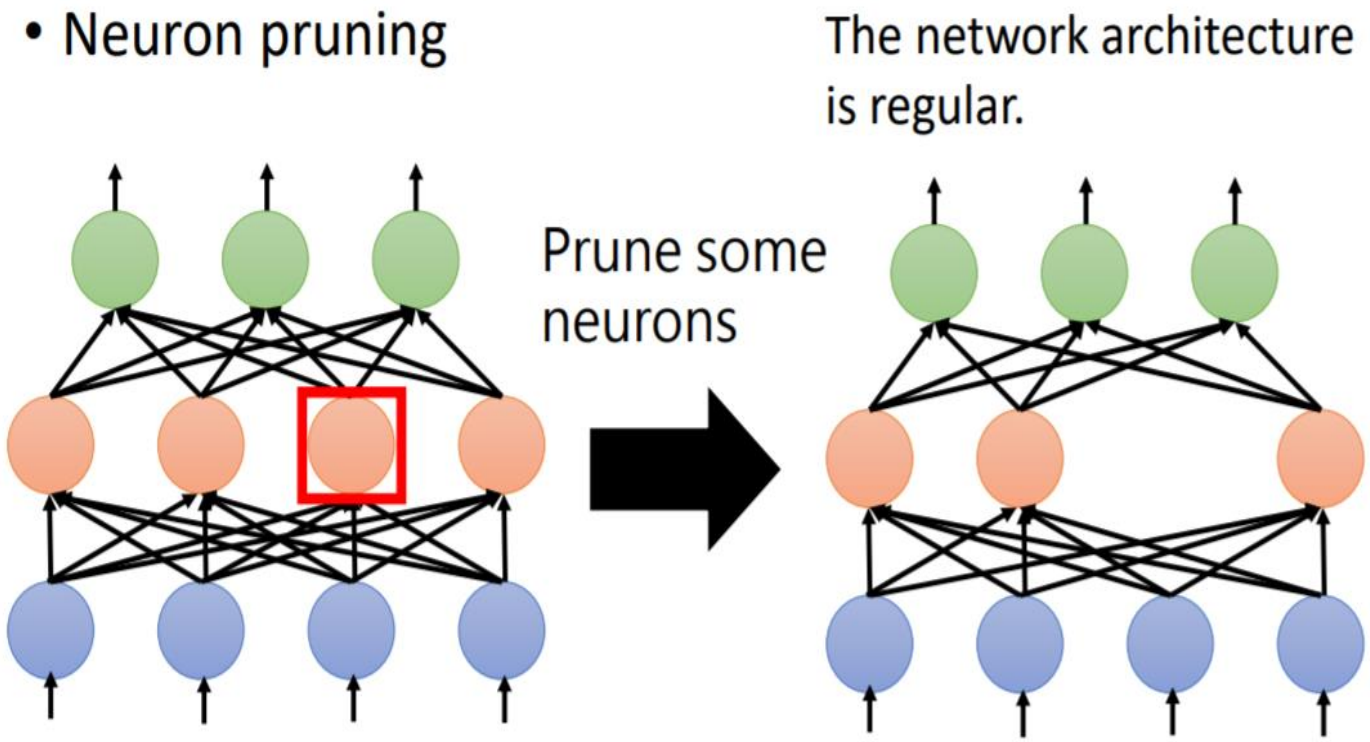
知识蒸馏 Knowledge Distillation

参照训练出来的网络，重新训练一个规模较小的网络

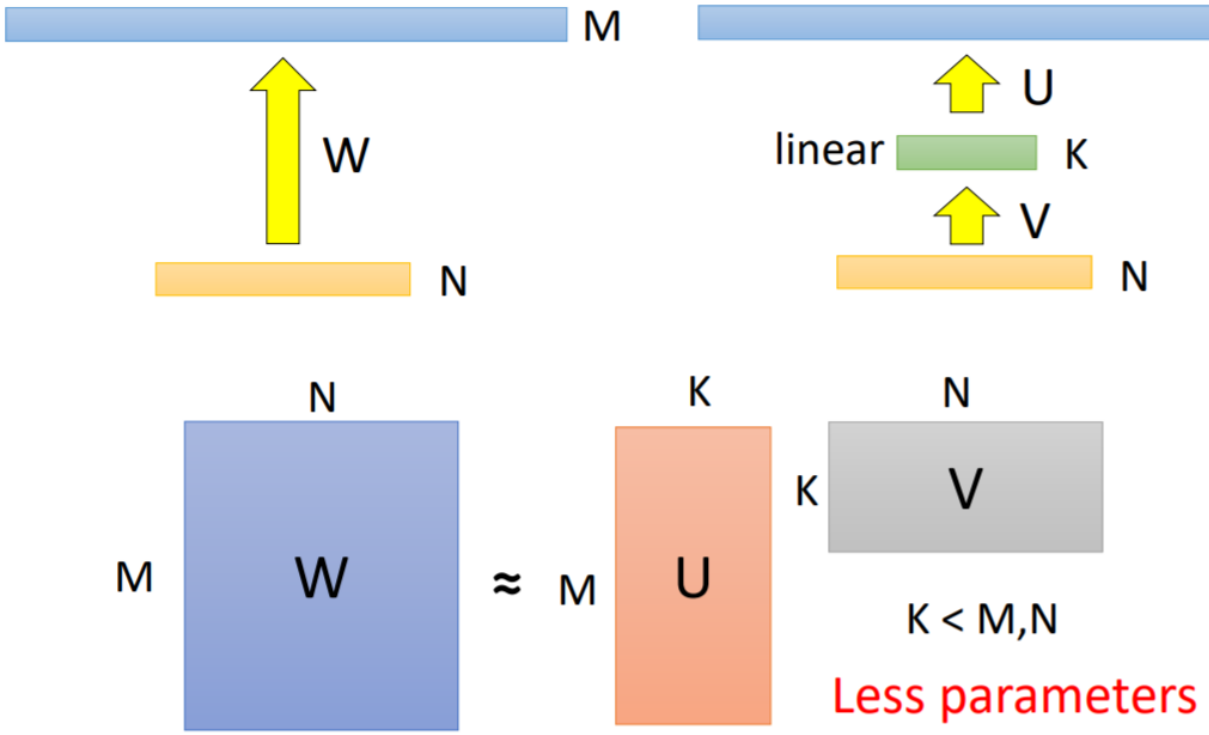


网络剪枝 Network Pruning

- Neuron pruning



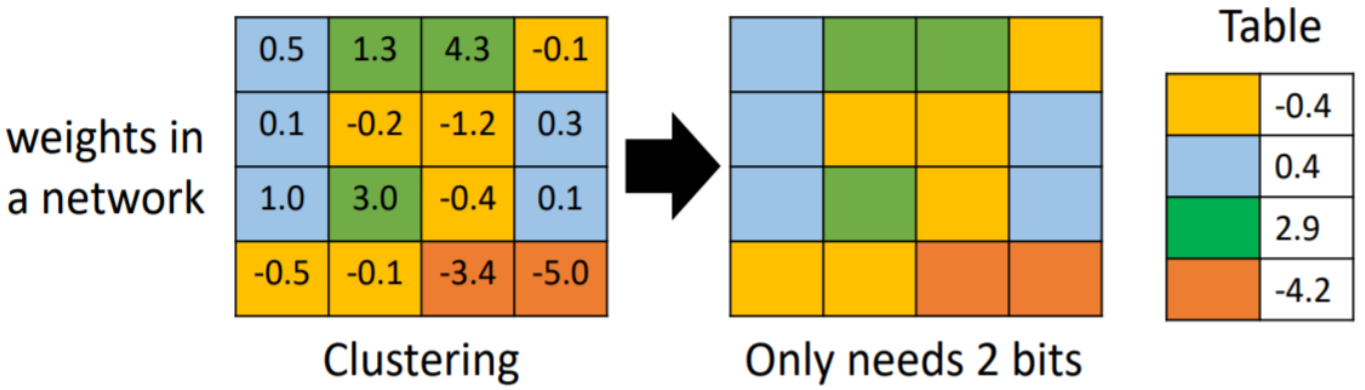
用少量参数来做 CNN Architecture Design



- 输入N个结点，输出M个结点，参数量为$M \times N$
- 在其中插入一个结点数为K的隐藏层，使得参数量为$N \times K + M \times K$

参数量化 Weight Quantization

- 1. Using less bits to represent a value
- 2. Weight clustering



- 3. Represent frequent clusters by less bits, represent rare clusters by more bits
  - e.g. Huffman encoding

- 将值相近的参数划分为一类，同时计算每一类参数的平均值
- 可以使用哈夫曼编码进一步缩减参数空间占用

Master Thesis Report

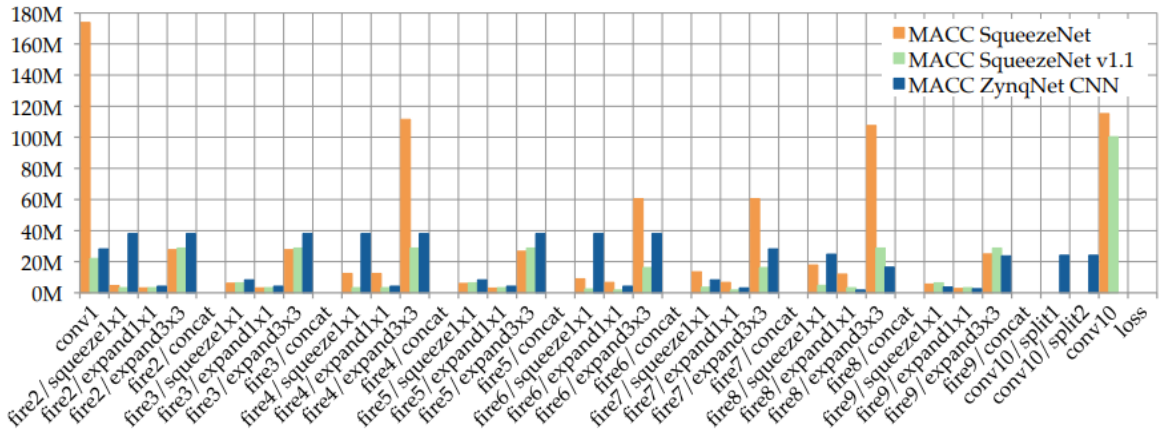
ZynqNet:

An FPGA-Accelerated Embedded Convolutional Neural Network

David Gschwend
davidgs@student.ethz.ch

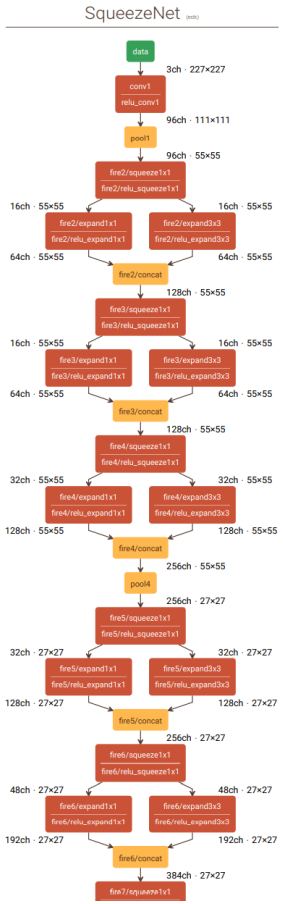
论文

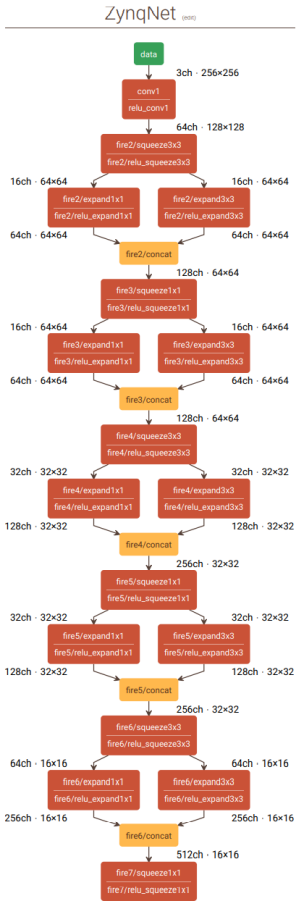
ZYNQNET



Optimizations

1. 基于SQUEEZE NET。调整结构，减少了整体参数量
2. 减少了CONV10中不必要的PADDING，使得结果中MACC减少了30%





Optimizations

1. 维度全部为2的指数
2. 全卷积网络，没有Max-Pooling