Compression

Yu Xueqing

2022 年 8 月 8 日

1 Network Pruning

Network pruning: 把冗余的参数删去
pre-trained network -> evaluate the importance -> prune, accuracy drop -> fine-tune -> evaluate ->...

- weight pruning 以参数为单位修剪，修剪后网络不规则，训练不好加速
- neuron pruning 以神经元为单位修剪

先训练小的网络，再逐渐扩大？large network is easier to optimize
Lottery Ticket Hypothesis: 有几率 roll 到更好的 sub-network 的初始值，prune 留下来的是好的 initial weight 训练出来的 sub-network

2 Knowledge Distillation

student net(small) 向 teacher net(large) 学习，cross-entropy minimization，比用绝对的 0/1 label 好
Ensemble: 多个模型结果平均
Temperature for softmax: $y_i' = \frac{e^{y_i}}{\sum e^{y_j}}$ -> $y_i' = \frac{e^{y_i/T}}{\sum e^{y_j/T}}$ ，均匀分布. T 是 hyperparameter

3 Parameter Quantization

1. using less bits to represent a value
2. weight clustering 参数分类，一个类内所有参数相同
3. 根据频率用不同长短编码表示，e.g. Huffman encoding

binary weights

4 Architecture Design

depthwise separable CNN

1. Depthwise Convolution：input 几个 channel，就有几个 filter，每个 filter 卷一个 input channel

2. Pointwise Convolution: 都是 1 x 1 filter, 个数等于 output channel, 在 depthwise conv 的结果上做普通卷积

depthwise+pointwise convolution 比普通 CNN 总参数量少。相当于 low rank approximation

5 Dynamic Computation

network 自由调整需要的运算量
dynamic depth: 取中间层的输出
dynamic width: 取每一层的部分 neuron