Network pruning
网络剪枝

因为神经网络中是过参数的，很多参数没有用

如何评估一个参数重要与否？
如果绝对值很小，则说明不重要

如何评估一个Neuron是否重要？
给定dataset，如果一个neuron的输出基本全为0，则说明不重要

①有一个训练大网络
②评估重要性
③删掉不重要
④微调 Fine-tune
⑤循环 ②-④ 直到满意

按照重要性排序，删掉不重要的

Fine-tune

Q：为什么不直接train小模型，而用大模型蒸馏？
A：小模型难train，大模型容易优化

如果裁剪参数weight的话，矩阵运算可能不规则（将weight设为0）
所以主要裁剪 neuron

知识蒸馏：
Student Net 模仿 Teacher Net 的输出

用cross-entropy来minimize输出之间的loss
distribution

Teacher Net 教给学生的更多
例如：在识别数字1时，有输出 confidence

1：0.7　9：0.1
7：0.2

教给学生时会告诉 Student Net 为 1 和 7 长得像.

同时 knowledge distillation 也可以进行 ensemble Learning 的操作.

知识蒸馏的小 trick.

Temperature 方式:

$y_i = softmax(\vec{x})$ $\quad y_i = softmax(\vec{x}/T)$

平滑化,避免高 confidence 忽略各个低 confidence 之间的相似性

参数量化: ①用更少的位来表征一个值

②权重聚簇.

例如:

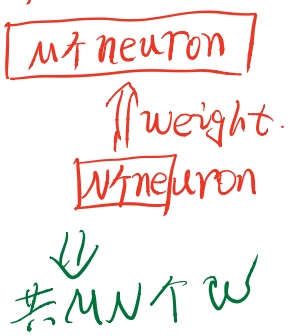
1	2	2	1
3	2	1	4
1	4	2	3
4	2	4	3

1, 2, 3, 4 代表 4 种不同范围的权重, 如 [0.1) [2, 3)

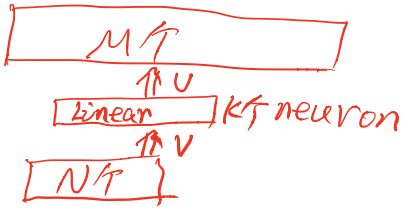
存储位表 将位置处的表值改为类别编码(可为 huffman 编码)

然后存储每种类别的均值(或其他方式)即可

架构设计：



共MN个w



共 $KN+KM$ 个w

其中 $K<M,N$