YOUR NAME 1/23/18

ZhengZishuo

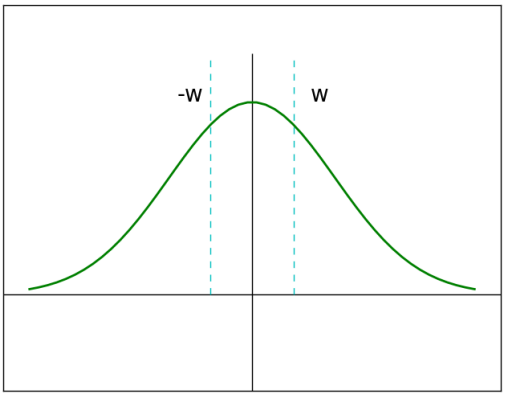
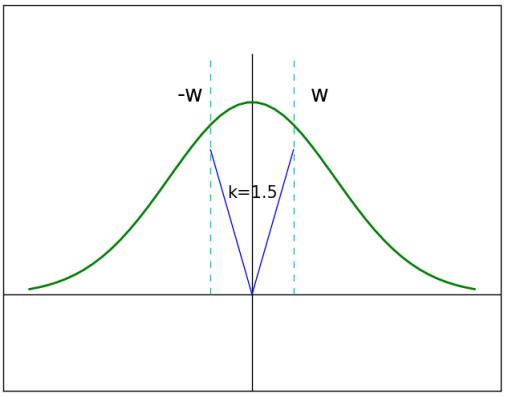
WEEKLY STATUS REPORT

Period Ending: 2017.10.17

ACTIVITIES COMPLETED THIS WEEK

神经网络压缩：

参照 ICLR 2016 会议论文《Deep Compression: compressing Deep Neural Networks with pruning, Trained quantization and Huffman coding》 所提出的压缩步骤中的第一步，剪枝，进行剪枝阈值的调整以及用概率剪枝函数代替决定性剪枝函数的尝试。实验在 cifar-10 上进行。

决定性剪枝	概率剪枝
	

决定性剪枝：

选定剪枝阈值 w 后，对于神经网络中的参数值进行调整。

x = x , x >= w

x = 0 , x < w

概率剪枝：

由于决定性剪枝不够柔和，所以使用概率剪枝。

选定剪枝阈值 w，以及概率参数 k

P(x) = 1 , x >= w

P(x) = min(kx,1) , x < w

用 P(x)代表神经网络参数被保存下来的概率。

Algorithm

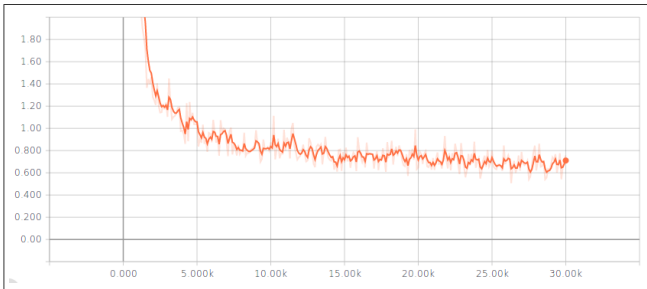
整体流程

1. 用 cifar10 的训练集对于神经网络进行训练，每 3k 步进行一次剪枝。
2. 按照合适的参数剪枝之后，装载剪枝后的参数，继续进行训练。
3. 总共进行 30k 步的训练。

剪枝流程

1. 设定超参数 准确性因子 accuracy_factor ， 压缩率因子 compress_factor，阈值搜索步长 dw。
2. 记录初始准确率 accuracy0 和 压缩率 compress0 。
3. 设置 w’ = 0
4. w’ += dw
5. 以 w’为阈值进行剪枝，载入参数后在测试集上进行测试，得到准确率 accuracy_t 与压缩率 compress_t。
6. 若 accuracy_factor * (accuracy0 – accuracy_t) + compress_factor * (compress0-compress_t) < 0
   则终止剪枝，此时的 w’即为最优阈值。
   保存剪枝结果，进行下一轮训练。
7. 跳转到第 4 步。（由于 w’单调递增，所以不用恢复神经网络参数）

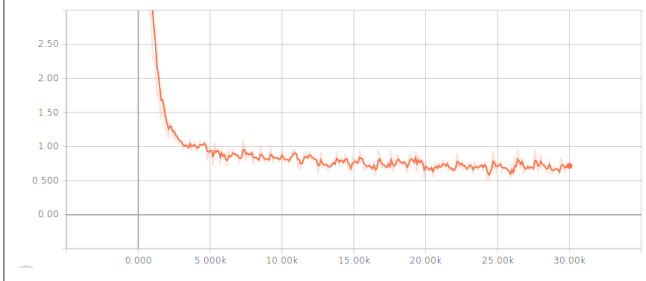
实验结果：loss 图

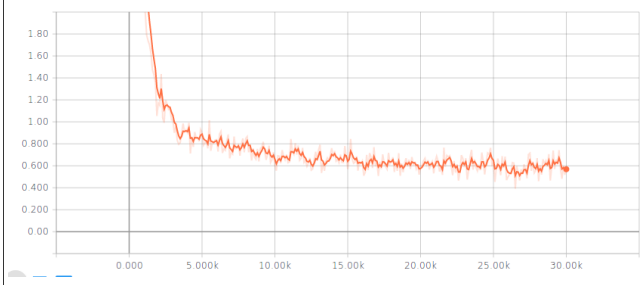


NoPruning

Accuracy Rate = 83.4 %

Compress Rate = 100 %





Pruning – accuracy:compress = 49:1

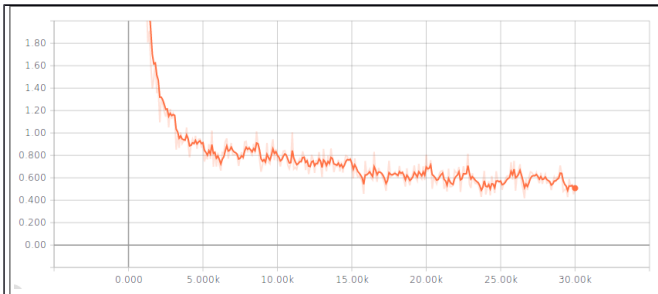
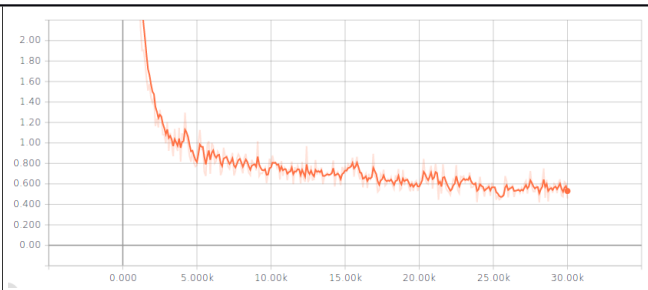
Accuracy Rate = 82.2%

Compress Rate = 3.25%

soft_Pruning(probability) – accuracy:compress = 49:1

Accuracy Rate = 82.3%

Compress Rate = 2.77%

	
Pruning – accuracy:compress = 499:1 **Accuracy Rate = 83.7%** **Compress Rate = 4.71%**	**soft_Pruning(probability) – accuracy:compress = 499:1** **Accuracy Rate = 83.5%** **Compress Rate = 4.74%**

IDEAS

- 在更大，更复杂，可以实际应用的数据集上进行实验。
- 这个 pruning 过程很像人脑的遗忘机制，结合更多的神经科学的知识进行改进。