Deep Net Triage: Analyzing the Importance of Network Layers via Structural Compression

Theodore Nowak[1] and Jason J. Corso[2]
Robotics Institute[1,2] and Electrical Engineering and Computer Science[2], University of Michigan, Ann Arbor, MI

Abstract

- Investigation of model quality after model compression, and varying initialization and training regimes
- 4 data sets of varying complexity, 124 experimental network combinations
- No layer is more contributory to overall accuracy
- Accuracy can be improved by a compressed network
- Knowledge Distillation caps accuracy at that of the parent

Introduction

- Deep network models are poorly understood
- Model pruning and compression has been shown to reduce overhead and maintain accuracy of networks [2]
- Knowledge Distillation has been used to approximate model ensembles and deeper models [3]
- How does the model cope with parameter loss, and what is the optimal model structure?

Problem Statement

Given the VGG16 network:

- How do the filters learned, the accuracy, and the rate of convergence change as groups of layers are replaced by a single layer?
- How does varying initialization and training regimes affect this?
- How does this vary across data set complexity?

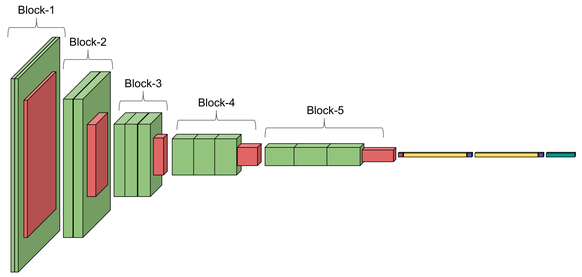
Experimental Overview

- Every combination of experiments was performed in addition to the 4 base models

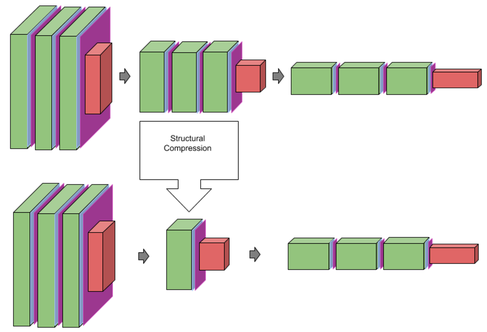
Experiment Sets	
Data Sets	{ MNIST, CIFAR10, CIFAR100, Stanford Dogs }
Layer Blocks Compressed	{ Block-1, Block-2, Block-3, Block-4, Block-5 }
Initialization Schemes	{ Random, Mean, Student-Teacher Network }
Training Regimes	{ Frozen Model, Thawed Model }

Method: Structural Compression

The VGG16 network structure consists of 5 repeated blocks of convolution layers followed by max pooling.



Structural Compression reduces one of these blocks of 2 to 3 layers down to a single layer, and reinitializes its weights.



Method: Experimental Methods

Mean Initialization

Initialize the compressed layer as an average of the layers it replaces.

$$f_{avg} = \frac{1}{N}\sum_{i=1}^{N} f_i \;\; where \;\; f_i \in R^{3x3x1}$$

Student-Teacher Network

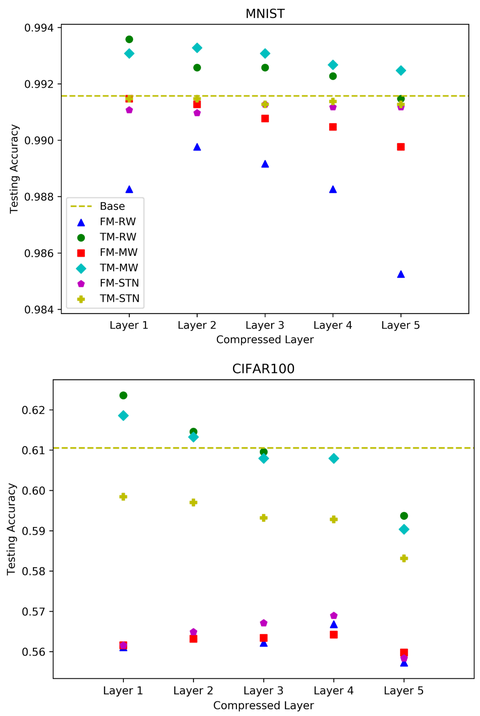
Compute and backpropagate the loss after the compressed layer.

$$L_{STN} = \min_{W_s,b_s} \frac{1}{N}\sum_{i=1}^{N} \|s(x_i, W_s, b_s) - t(x_i, W_t, b_t)\|^2$$

Frozen vs. Thawed Model

We test whether the compressed layer can learn filters to fit into the existing model. The frozen model freezes the rest of the model parameters, while the thawed model allows for gradient updates across the entire model.

Results: Accuracy



Maximal accuracy achieved by each model on each data set for each compressed layer.

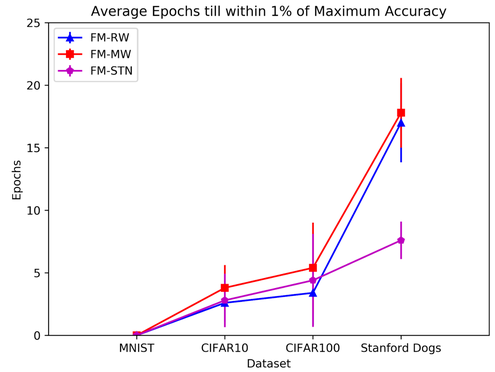
- No layer is more contributory
- Accuracy can be improved by compression
- Knowledge Distillation caps accuracy -- Why?

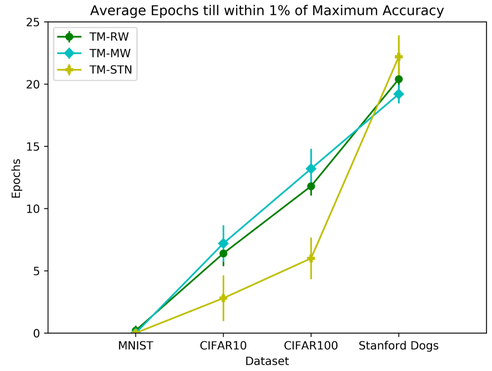
Key Acronyms

Base: Uncompressed
FM: Frozen Model
TM: Thawed Model
RW: Random Weights
MW: Mean Weights
STN: Student-Teacher Network

Results: Rate of Convergence

Relative to accuracy, STN networks tend to converge more quickly, while otherwise, convergence rate and relative accuracy are directly correlated.

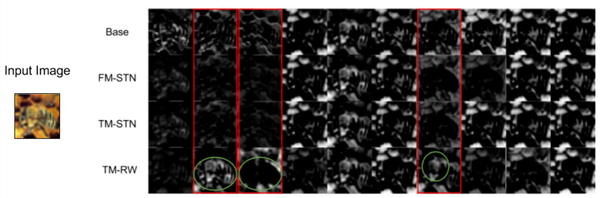




Results: Qualitative

Q: Why does knowledge distillation cap the accuracy at that of the base model?

A: The Student network is encouraged to learn the same filters as the Parent rather than different, more efficient filters.



References

[1] Nowak and Corso. *Deep Net Triage: Analyzing the Importance of Network Layers via Structural Compression. ArXiv 2018.*

[2] Han et al. *Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding.* ICLR 2015.

[3] Hinton et al. *Distilling the Knowledge in a Neural Network.* ICLR 2015.