Representation Aware Pruning with Centered Kernel Alignment

Calvin Higgins

Department of Computer Science and Statistics
Department of Mathematics and Applied Mathematical Sciences
University of Rhode Island

CSC 561 Project Presentation, May 2023

Table of Contents

Introduction

Neural networks are **big**!

- High memory footprint
- Compute intensive inference

Can we shrink them with harming accuracy?

- Reduce memory consumption
- Increase inference throughput

Table of Contents

Structured Pruning

Remove neurons from the network

- Reduces memory consumption
- Increases inference throughput

Need to minimize damage to the activations!

How can we measure this?

Centered Kernel Alignment (CKA)

Measure of similarity between layer activations

- 1 is complete similarity
- 0 is no similarity

Table of Contents

CKA Pruning

1. Compute activations with 512 samples.
2. For each layer
    1. For each neuron
        1. Zero neuron weights and recompute activations.
        2. Compute CKA between original and new activations.
        3. Restore neuron weights.
    2. Prune neuron whose removal resulted in maximum CKA.
    3. Repeat until $p\%$ of neurons are pruned.
3. Re-train network.
4. Repeat j times.

L1 Pruning

1. For each layer
   1. Compute L1 norm of neuron weights.
   2. Prune p% of neurons with lowest L1 norm.
2. Re-train network.
3. Repeat j times.

Dataset and Preprocessing

MNIST Dataset

- 28×28 images of handwritten digits
- 55000-5000-10000 train-validation-test split
- Normalized with 0 mean and 1 SD

Training

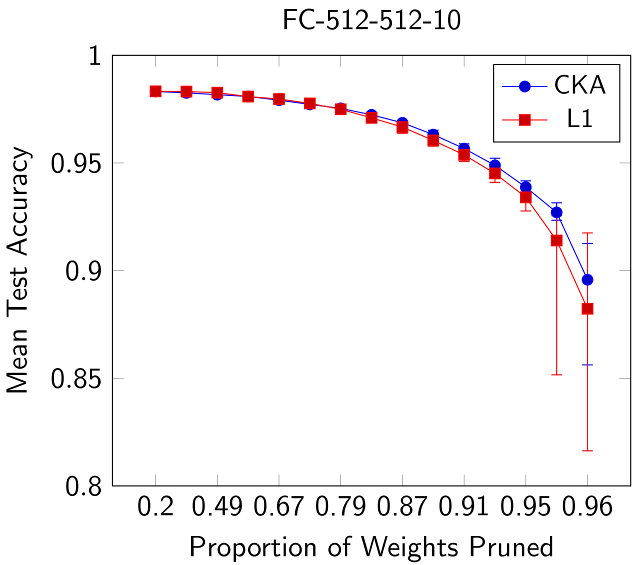
All models are MLPs trained with

- Adam with $\gamma = 0.001, \beta_1 = 0.9$ and $\beta_2 = 0.999$
- Maximum 50 epochs with ES on validation loss (3 epoch patience)
- 50% dropout on hidden layers
- Batch size of 512

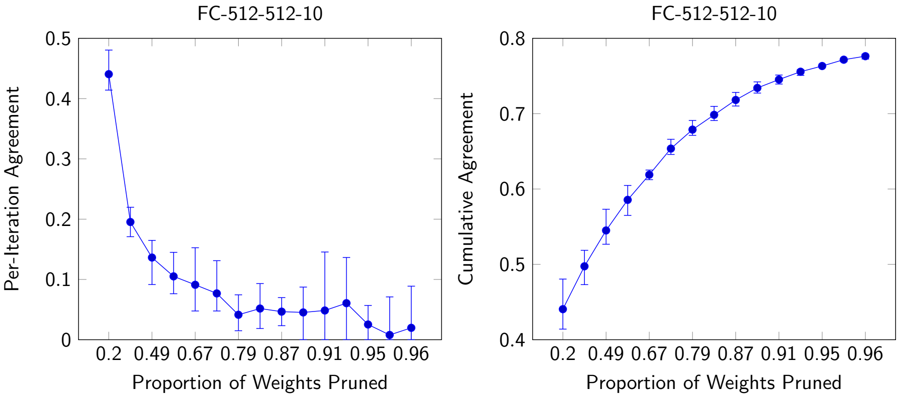
The learning rate was found via hyperparameter search with WandB.

Table of Contents

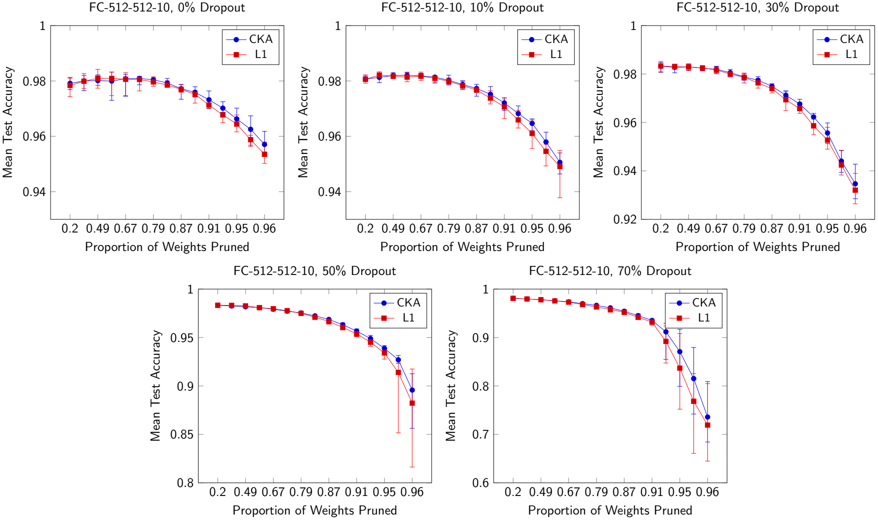
Iterative Pruning Accuracy



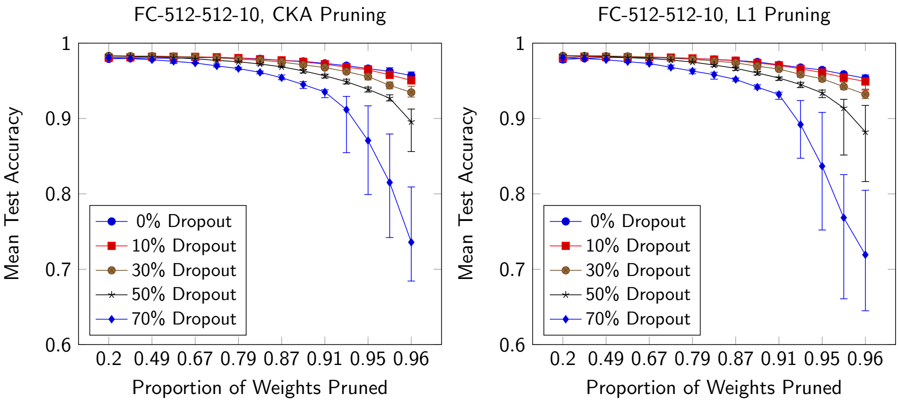
Iterative Pruning Agreement



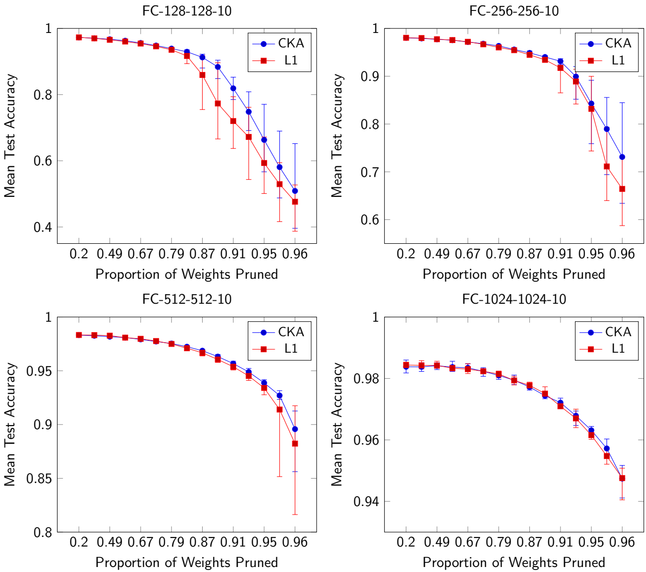
Varying Dropout



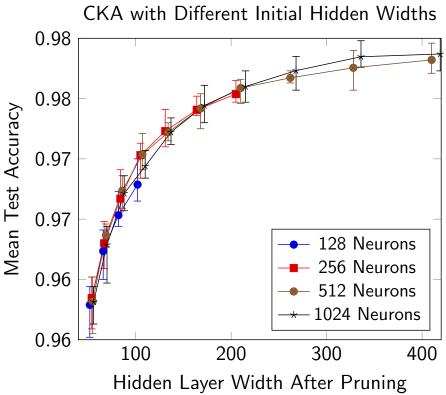
Varying Dropout

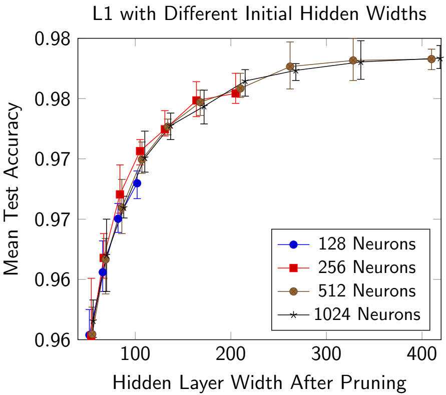


Varying Width

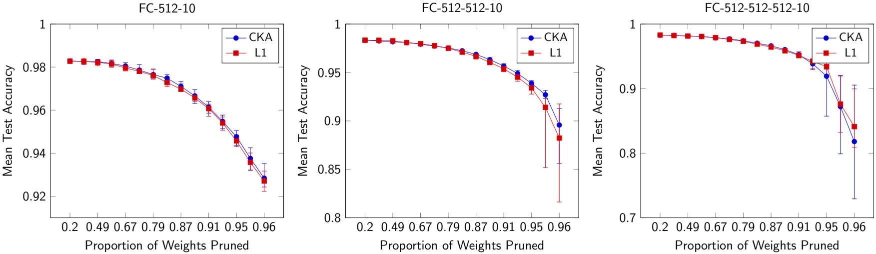


Varying Width





Varying Depth



Varying Pruning Rate

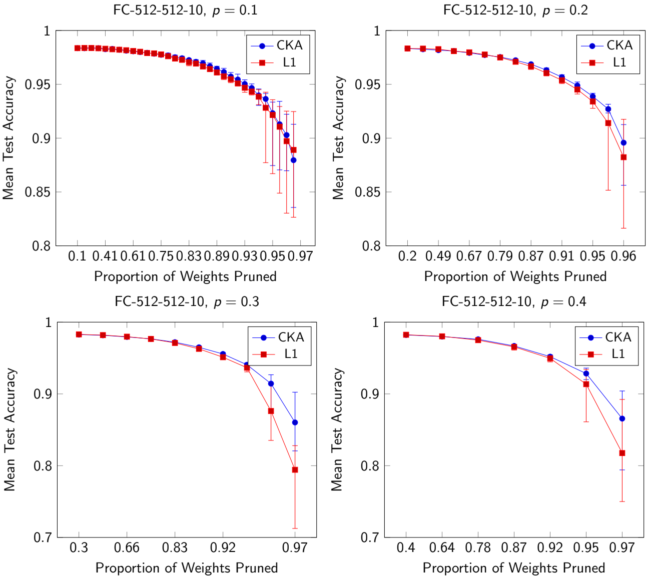


Table of Contents

Conclusions

CKA pruning offers marginal benefits

- At high pruning rates
- For heavily pruned networks

Rendered useless by time complexity...