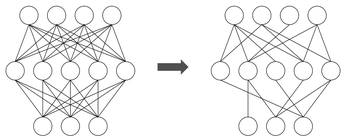
What is the State of Neural Network Pruning?

Davis Blalock*, Jose J. González*, Jonathan Frankle, John Guttag

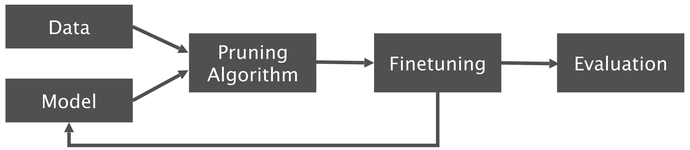
Summary

- Meta-analysis of neural network pruning papers
- Main finding: lack of experimental standardization has made it impossible to tell what works best
- Open source library for standardized evaluation of pruning methods: https://github.com/jjgo/shrinkbench

Neural Network Pruning

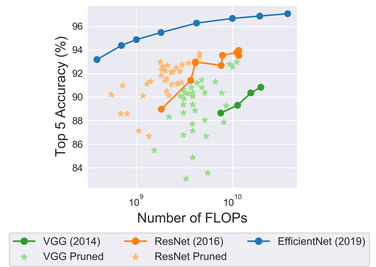


- Pruning: given a fixed architecture, remove parameters to increase efficiency
- Typical pipeline: iteratively prune and finetune a pretained model



Consistent Findings

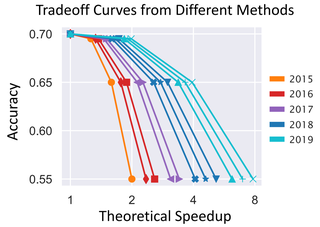
- Pruning works!
  - Measurable efficiency gains
  - Better than random
- But not as helpful as using a better architecture



Lack of Experimental Standardization

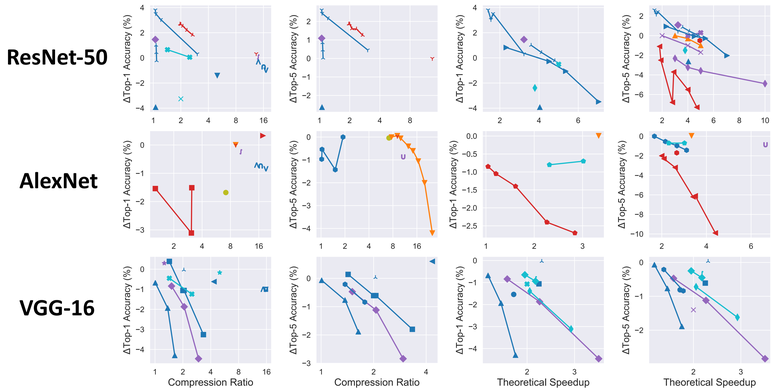
Ideal

- Same dataset, architecture, efficiency metric, quality metric
- Whole tradeoff curve reported
- Identical hyperparameters
- Consistent progress over time

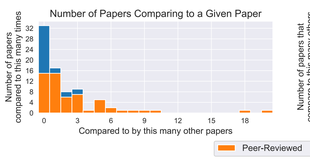


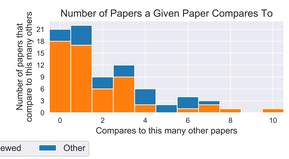
Reality

- Fragmented datasets, architectures, metrics
- Different hyperparameters
- Incomplete curves, or single points
- No clear progress over time



But Why?

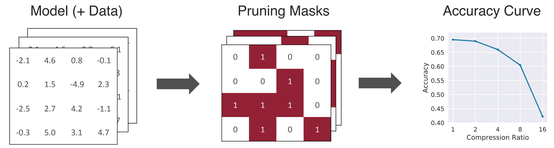




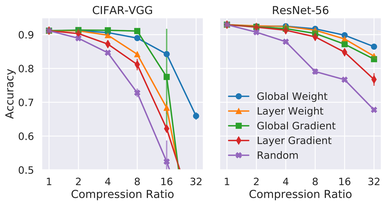
Dataset	Architecture	# of Papers
ImageNet	VGG-16	22
CIFAR-10	ResNet-56	14
ImageNet	ResNet-50	14
ImageNet	CaffeNet	11
ImageNet	AlexNet	9
CIFAR-10	CIFAR-VGG	8
ImageNet	ResNet-34	6
ImageNet	ResNet-18	6
CIFAR-10	ResNet-110	5

ShrinkBench

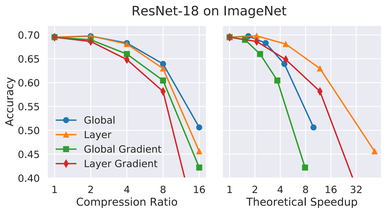
- Standardized datasets, architectures, preprocessing, finetuning, evaluation, etc
- Flexible mask-based API to allow arbitrary sparsity and pruning algorithm



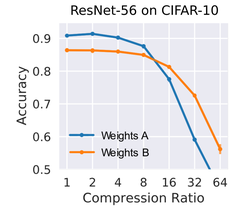
Additional Pitfalls Uncovered



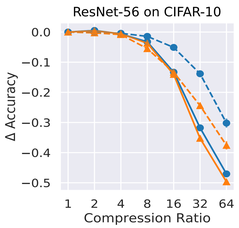
- Must use multiple networks and amounts of pruning



- Size and speed metrics are not interchangeable



- Initial weights are an important confounder



- Reporting changes in accuracy doesn't fix it

This research was funded by two Qualcomm Innovation Fellowships, a "La Caixa" Fellowship, Wistron Corporation, and Quanta Computer

{dblalock, josejg}@mit.edu