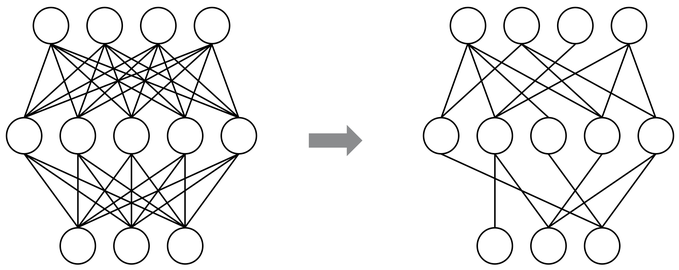
Standardizing Neural Network Pruning

Jose Javier Gonzalez Ortiz, Davis Blalock, John Guttag

Neural Network Pruning

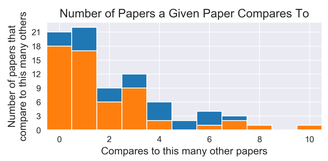
Pruning: Systematically removing parameters from an existing network

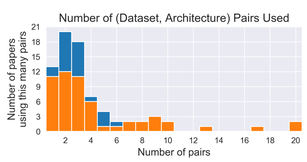
Goal: Reduce size of network as much as possible with minimal drop in accuracy

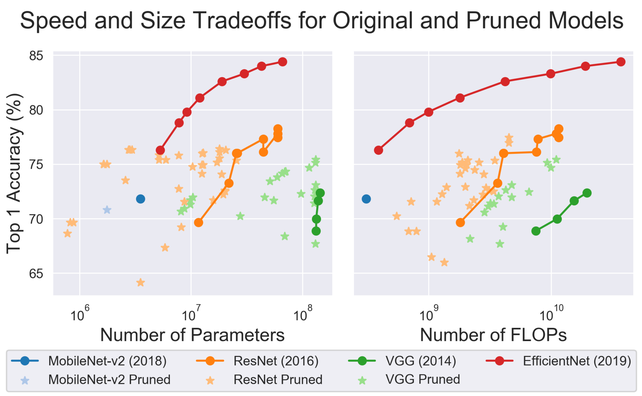


A survey of 80 pruning papers revealed

- No clear standardized metrics or baselines
- No clear state-of-art due to the lack of standardized evaluation





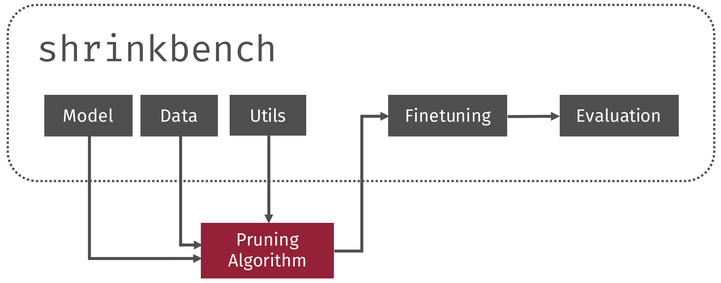


We introduce ShrinkBench, a tool for standardizing NN pruning evaluation.

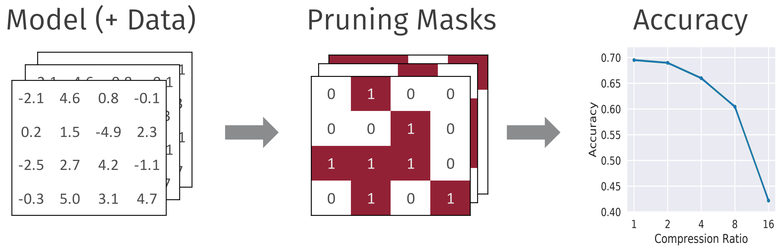
`shrinkbench.github.io`

ShrinkBench

- Open-source library to facilitate standardized neural network pruning evaluation
- Provides standardized datasets, pretrained models, and evaluation metrics



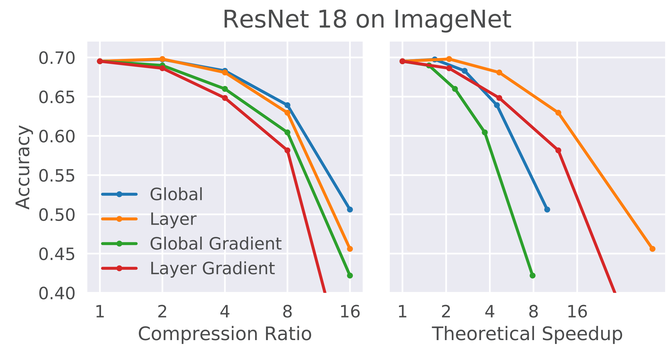
- Enables rapid prototyping and evaluation of pruning methods
- Simple and generic parameter masking API
- Measures number of nonzero parameters, activations, and FLOPs



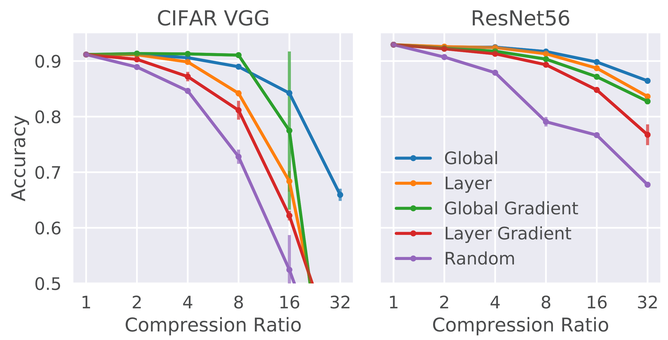
- Empirical results using pruning baselines show the need for standardized evaluation.

Results

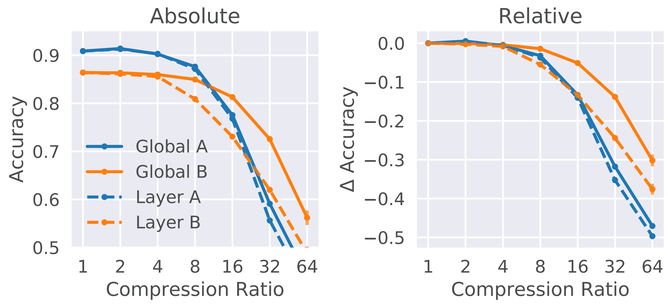
1. SB returns both compression & speedup since they interact differently with pruning



2. SB evaluates with varying compression and with several (dataset, architecture) combinations



3. SB controls for confounding factors such as pre-trained weights or finetuning schedules



This research was funded by "la Caixa" Foundation Fellowship and by Qualcomm Innovation Fellowship

josejg@mit.edu