Explainable Pruning for Deep Neural Networks Using Reinforcement Learning

Jeremy Lee, Saranya Vijayakumar, Alex Gaudio, Christos Faloutsos, Asim Smailagic, and Mahadev Satyanarayanan
Carnegie Mellon University

Neural Networks Are Too Big for Mobile Devices

How do we make deep models mobile friendly?

•Deep Neural Networks (DNN) are cumbersome, and prohibitive to be adopted on mobile devices. They might:

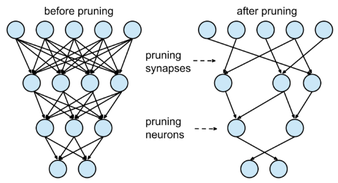
1. Increase the power consumption
2. Occupy memory and storage
3. Result in a long inference time

•Given a well-performing pre-trained deep model, our goal is to:

1. **Identify** the importance of parameters
2. **Prune** the model structurally, removing least informative channels
3. **Provide** explanations to the pruning

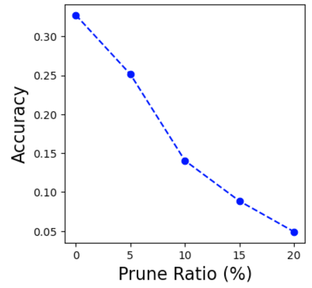
•**Challenges**

1. Deep models are not interpretable because of the non-linear transformations.
2. Most of the pruning strategies are performance-oriented and therefore hard to explain.



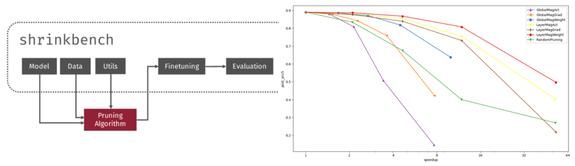
Preliminary & Benchmark

Naïve channel pruning is not sufficient to maintain accuracy. As an example, we use a subset of 200 classes of ImageNet. We load the weights of a pre-trained ResNet50 model (for 1000 classes).
We remove the channels with lower weights from each layer. Performance degrades rapidly.
Retrain is needed.



Benchmark

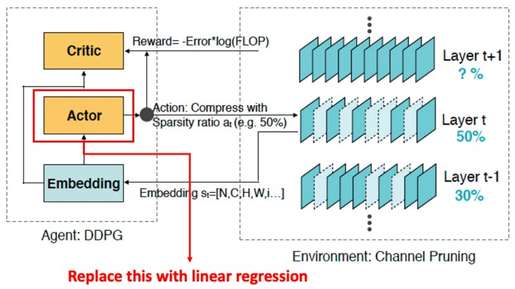
To compared with other competitors, we use ShrinkBench for a more comprehensive comparison.



The theoretical speedup of a model via different pruning techniques

Our Approach

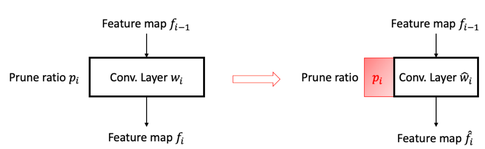
Architecture



Features

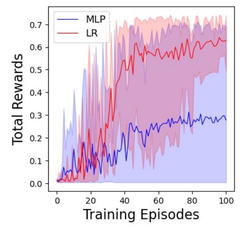
•**Layer Index:** The smaller the shallower in the deep model
•**Largest Singular Value**
•**Effective Rank:** 90% of energy
•**Is Conv:** 1 for yes and 0 for no
•**Is FC:** 1 for yes and 0 for no
•**In Channels:** # of channels in the input image
•**Reduced FLOPs:** Total # of reduced FLOPs in previous layers
•**Rest FLOPs:** Total # of remaining FLOPs in following layers
•**Last Action:** The pruned ratio to the previous layer

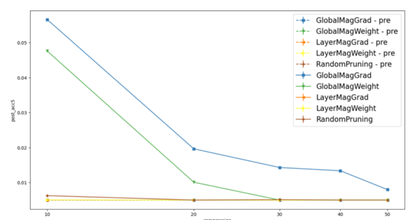
Pseudo Pruning



Performance and Results

•Our goal is to prune MobileNets.
•LR converges faster than MLP and has higher rewards.
•LR turns out to be a better choice over MLP, while being interpretable.
•50% of the FLOPs is pruned and the accuracy remains at 73.4%.





•However, every pruning strategy severely distorts the accuracy, which is hard to recover using only 1 epoch of retraining.

•One discovery is that the feature importance is not consistent across different episodes.
•To address that, we select the feature with L1 regularization, where the feature importance is now much more stable.