Network Pruning that Matters: A Case Study on Retraining Variants

ICLR 2021

Duong H. Le, Binh-Son Hua

VinAI RESEARCH

VINUNIVERSITY

Motivation

Lin et al., 2020: 30 epochs *each iteration*, learning rate initialized at 0.01, drop by factor of 10 after 5 epochs.

Lin et al., 2019: 100 epochs, learning rate fixed at 0.001.

CIFAR-10

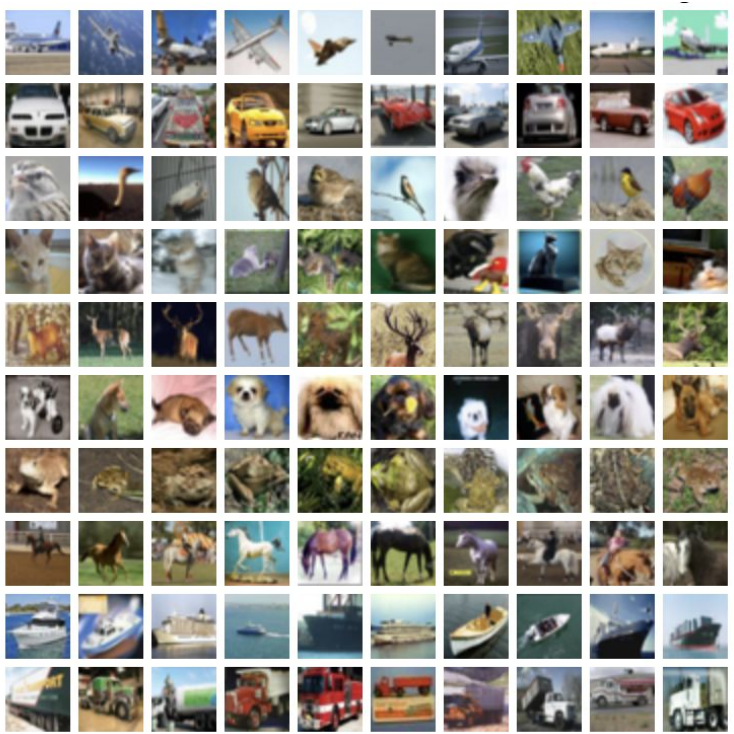


Image source: https://www.cs.toronto.edu/~kriz/cifar.html

Motivation

Li et al., 2020: 120 epochs, learning rate initialized at 0.01 drop by factor of 10 after 30% and 60% of training

Lin et al., 2020: 30 epochs *each iteration*, learning rate initialized at 0.01, drop by factor of 10 after 10 epochs.

Aflalo et al., 2020: 50 epochs, *doesn't specify learning rate schedule*, employ knowledge distillation.

He et al., 2019: 90 epochs, learning rate initialized at 0.01, drop by factor of 10 after 30 epochs.

Lin et al., 2019: 30 epochs, learning rate initialized at 0.01, drop by factor of 10 after 10 epochs.

ImageNet

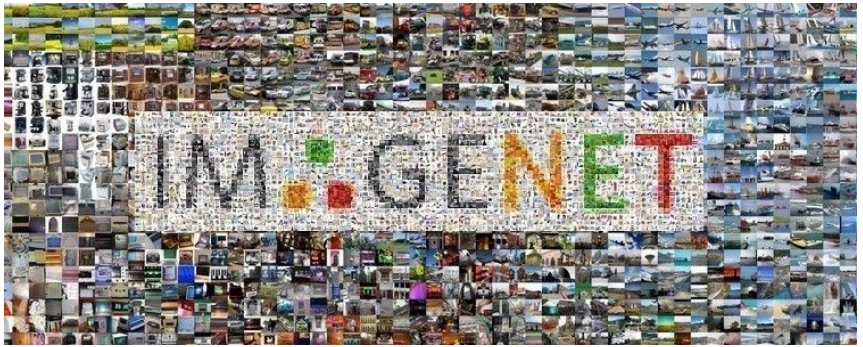


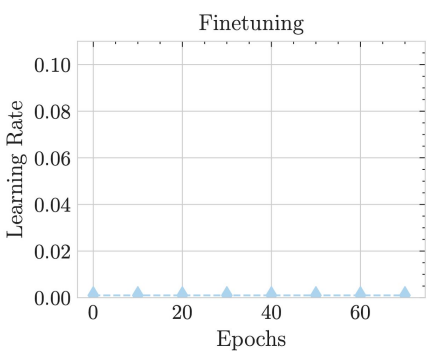
Image source: https://cv.gluon.ai/build/examples_datasets/imagenet.html

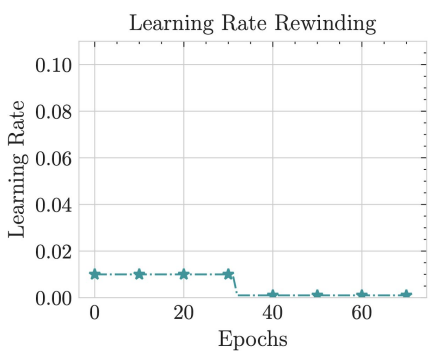
Case Study on Retraining Variants

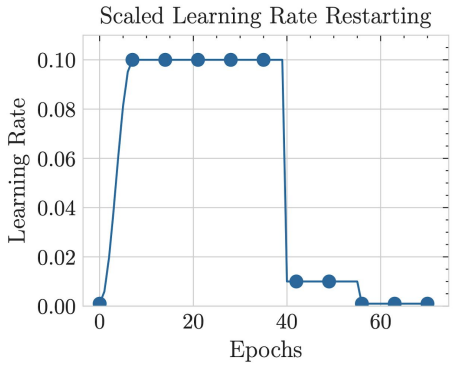
Variants of Learning Rate Schedules

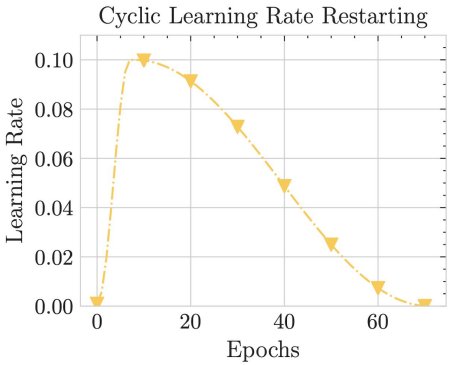
- **Fine-tuning (FT):** continue train the pruned networks for *T* epochs with the last (smallest) learning rate of original training.
- **Learning Rate Rewinding (LRW):** When retraining for *T* epochs, instead of using fixed learning rate (fine-tuning), learning rate rewinding employ the learning rate from previous *T* epochs.
- **Scaled Learning Rate Restarting (SLR):** learning rate schedule that is proportionally identical to the standard training.
- **Cyclic Learning Rate Restarting (CLR):** leverage the 1-cycle (Smith & Topin, 2019), which is shown to give faster convergence speed than conventional approaches

Variants of Learning Rate Schedules







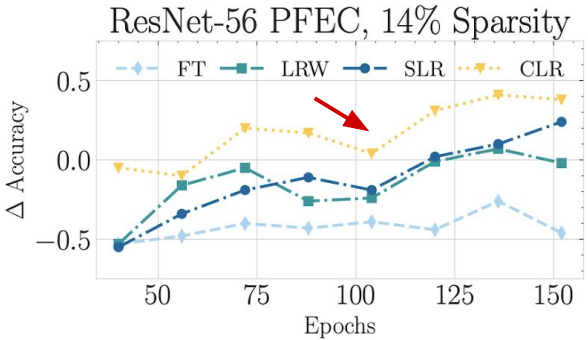


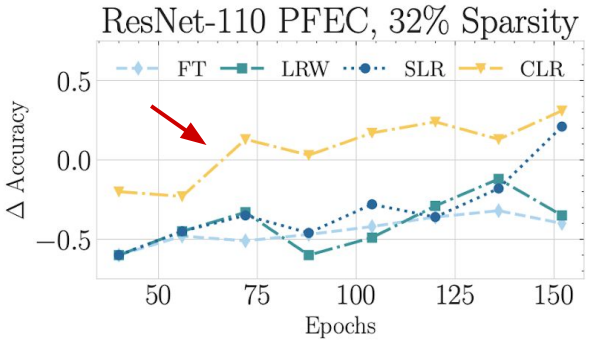
Example: Training the network for T = 160 epochs:

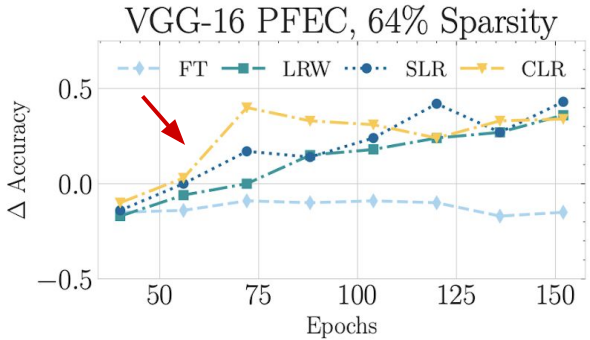
- Epoch 0 -> 80: lr=0.1
- Epoch 81 -> 120: lr=0.01
- Epoch 121 -> 160: lr=0.001

And retrain for **72** epochs!

Simple Settings







One-shot Structured Pruning on CIFAR-10 (Li et al., 2016)

Large learning rates (LRW, SLR, CLR) are significantly better than fine-tuning.

Other Pruning Algorithms

Model	Unpruned	Prune + FT	Prune + CLR	FLOPs ↓ %	Params ↓ %
Taylor-FO-BN-56%	76.15	71.69	**72.51**	67.2	66.8
Taylor-FO-BN-72%		74.50	**75.22**	45.0	44.5
Taylor-FO-BN-81%		75.48	**75.67**	35.0	30.1

Taylor Pruning on ImageNet (Molchanov et al., 2019)

CLR also improve the performance of more sophisticated pruning algorithms.

Pruning Algorithm vs Retraining

Pruning Algorithm: We use filters pruning (Li et al., 2016) for our simple baseline. The number of pruned filters in each layers is approximately the same so that the number of reduced parameters match with the target pruning algorithm (that are compared with).

Retraining schedule: We use CLR with the maximum learning rates are selected based on the heuristic proposed by Renda et al., (2020) i.e. learning rate rewinding. The minimum learning rate is equal 0.001*max_lr.

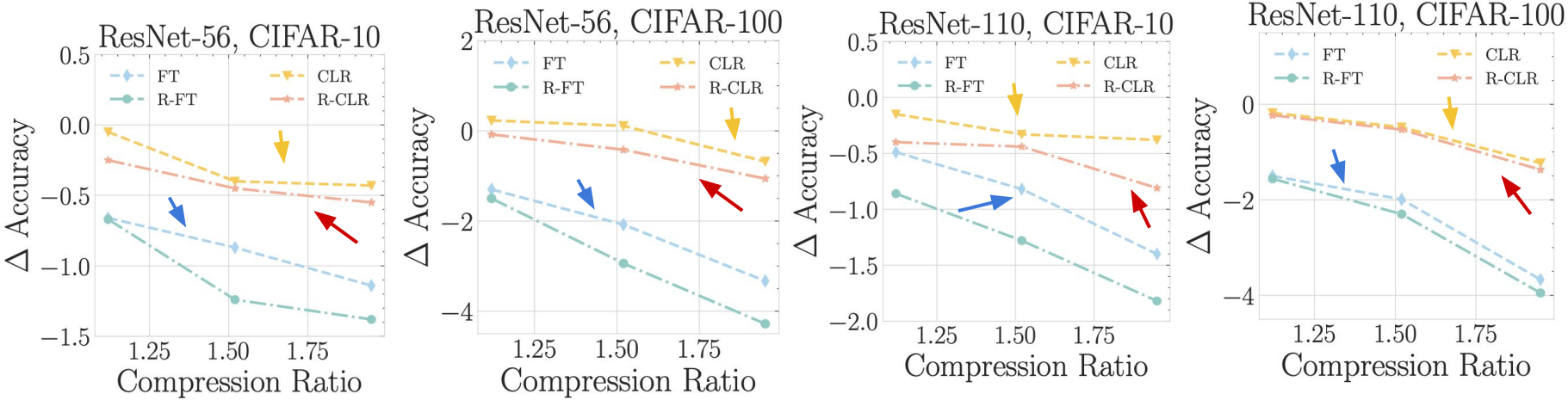
Discrimination-aware Channel Pruning

Model	Param ↓ %	FLOPs ↓ %	Method	Unpruned Top-1	Top-1
ResNet-18	28.1	27.1	DCP	69.64	69.21
	31.9	-	PFEC + CLR	69.76	**69.31 ± 0.06**
	47.1	46.1	DCP	69.64	67.35
	50.6	-	PFEC + CLR	69.76	**67.38**
	65.7	64.1	DCP	69.64	**64.12**
	-	-	PFEC + CLR	69.76	64.08
ResNet-50	33.3	35.7	DCP	76.01	**76.40**
	33.7	-	PFEC + CLR	76.15	76.03
	51.4	55.5	DCP	76.01	74.95
	51.5	-	PFEC + CLR	76.15	**75.16**
	65.9	71.1	DCP	76.01	72.75
	66.1	-	PFEC + CLR	76.15	**72.92**

Our simple baseline vs DCP on ImageNet (Zhuang et al., 2018)

Extreme Cases with Random Pruning

Random Pruning with Restarting



Filters Pruning on CIFAR-10 and CIFAR-100

Retraining matters and should be standardized for fair comparisons!

Thanks!

https://lehduong.github.io

v.duonglh5@vinai.io