RESNET-18 COMPRESSION FOR CIFAR-10

Structured Pruning and Post-Training Quantization

Technion

Presentation By Din Alon & Yakir Witkin

Spring 2025

MOTIVATION

- ResNet-18 is overparameterized for CIFAR-10

- Baseline 42.66 MB, 11M parameters

- 92.02% accuracy

- **Goal**: Reduce model size and inference time

TECHNIQUES OVERVIEW

- Automatic Mixed Precision (AMP)
- Structured Pruning
- Post-Training Quantization (INT8, FP16 via TensorRT)
- Combined Pruning + Quantization

BASELINE TRAINING

Trained ResNet-18 on CIFAR-10 for 100 epochs

Used SGD, cosine scheduler, weight decay

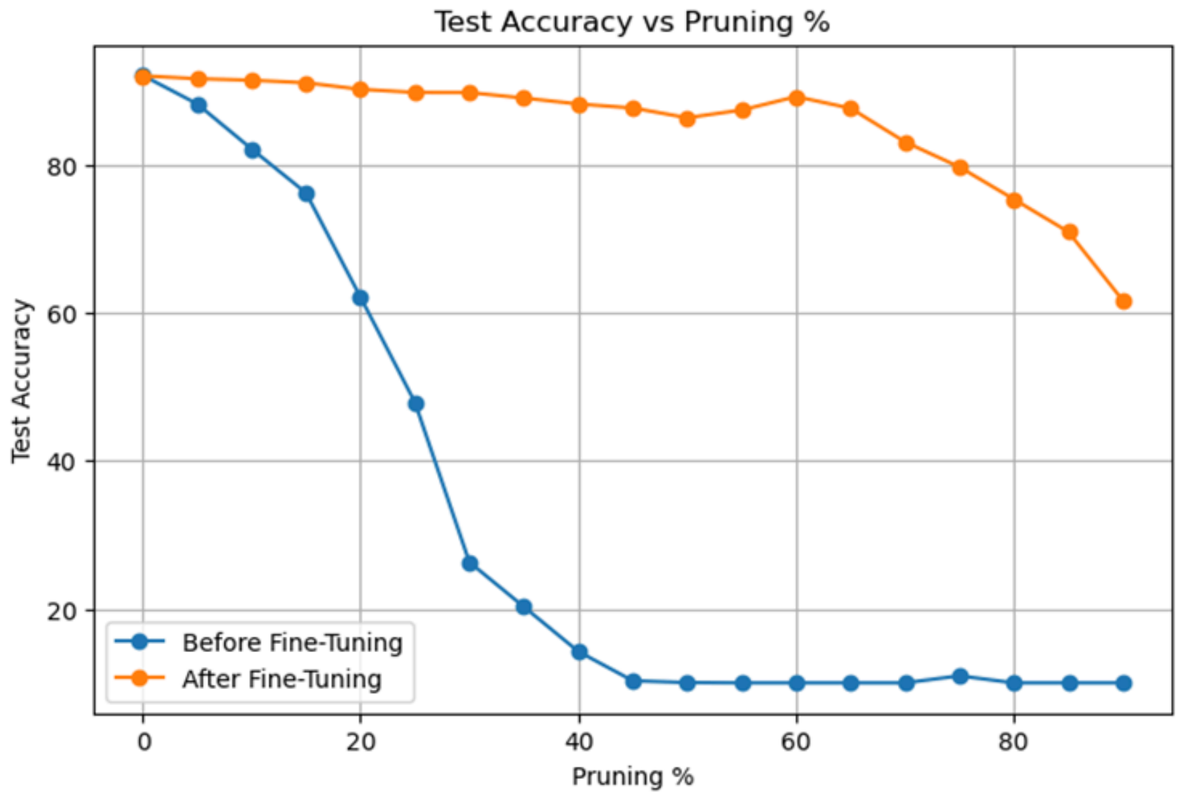
AMP TRAINING

- AMP via torch.cuda.amp
- Improves training speed 3.9 Vs 3.6
- Accuracy: 91.95% vs 92.02%

FP32

AMP

STRUCTURED PRUNING



- L1-norm filter ranking with torch-pruning
- Fine-tuned after pruning
- Best tradeoff: 60% pruning → 88.96% accuracy

STRUCTURED PRUNING – RETRAINING BEHAVIOR

More pruning → harder to recover accuracy levels and requires higher LR and more epochs to properly adapt.

Pruning %	Size (MB)	Parameters (M)	Inference Time (ms/image)	Accuracy Before Fine-Tuning (%)	Accuracy After Fine-Tuning (%)	Learning Rate	Epochs	Epoch Time (s)
0	42.66	11.17	1.82	92.02	91.95	0.00001	1	4.05
5	38.41	10.06	1.99	88.17	91.56	0.00005	3	3.93
10	34.44	9.02	2.02	82.09	91.36	0.00005	3	4.09
15	30.74	8.05	2.00	76.15	91.01	0.00005	3	3.82
20	27.20	7.12	2.21	62.09	90.13	0.00005	3	3.68
25	24.02	6.29	1.84	47.88	89.71	0.00005	3	3.40
30	20.85	5.46	1.94	26.30	89.69	0.00010	3	3.45
35	17.95	4.70	2.07	20.28	88.95	0.00010	3	3.47
40	15.32	4.01	2.04	14.24	88.15	0.00010	3	3.38
45	12.85	3.36	1.96	10.29	87.61	0.00010	3	3.55
50	10.69	2.80	1.90	10.04	86.29	0.00010	3	3.77
55	8.62	2.25	1.95	10.00	87.33	0.00010	10	3.77
60	6.79	1.78	1.90	10.00	89.14	0.00100	10	3.78
65	5.21	1.36	2.04	10.00	87.58	0.00100	10	3.91
70	3.82	1.00	1.83	10.00	83.03	0.01000	10	3.79
75	2.69	0.70	2.07	10.95	79.64	0.01000	10	3.86
80	1.70	0.44	1.82	10.00	75.29	0.01000	10	3.79
85	0.95	0.25	1.88	10.00	70.89	0.01000	10	3.78
90	0.42	0.11	1.73	10.00	61.63	0.01000	10	3.60

POST-TRAINING QUANTIZATION

CIFAR-10 Quantization Results Summary Table

Format	Model Size (MB)	Trainable Parameters (M)	Inference Time (ms/image)	Test Accuracy (%)
FP32 (TRT)	42.66	11.17	0.61	92.02
FP16	21.83	11.17	0.25	92.02
INT8	12.94	11.17	0.20	92.10

CIFAR-100 Quantization Results Summary Table

Format	Model Size (MB)	Trainable Parameters (M)	Inference Time (ms/image)	Test Accuracy (%)
FP32 (TRT)	42.84	11.22	0.61	79.26
FP16	21.91	11.22	0.25	79.22
INT8	12.96	11.22	0.19	79.07

- INT8 and FP16 via TensorRT
- Calibrated with class-balanced batches
- CIFAR-10 and CIFAR-100 tested

COMBINED PRUNING + QUANTIZATION

- Best result: 60% pruning + INT8
- Size: 2.96 MB, Accuracy: 89.03%

ResNet-18 Pruning + Quantization Summary

Pruning %	Format	Model Size (MB)	Trainable Parameters (M)	Inference Time (ms/image)	Test Accuracy (%)
50%	FP32	10.69	2.80	1.85	86.07
50%	FP16	5.97	2.80	0.17	86.06
50%	INT8	4.95	2.80	0.12	85.84
55%	FP32	8.62	2.25	2.04	87.28
55%	FP16	5.25	2.25	0.17	87.27
55%	INT8	3.44	2.25	0.13	87.24
60%	FP32	6.79	1.78	1.86	88.96
60%	FP16	4.40	1.78	0.17	88.95
60%	INT8	2.96	1.78	0.12	89.03
65%	FP32	5.21	1.36	2.06	86.85
65%	FP16	3.34	1.36	0.16	86.86
65%	INT8	2.31	1.36	0.12	86.70
70%	FP32	3.82	1.00	1.93	81.92
70%	FP16	2.61	1.00	0.16	81.93
70%	INT8	2.06	1.00	0.12	82.05
75%	FP32	2.69	0.70	1.83	80.04
75%	FP16	1.85	0.70	0.14	80.07
75%	INT8	3.13	0.70	0.11	80.09

FINAL RESULTS

Baseline

42.66 MB, 92.02% accuracy

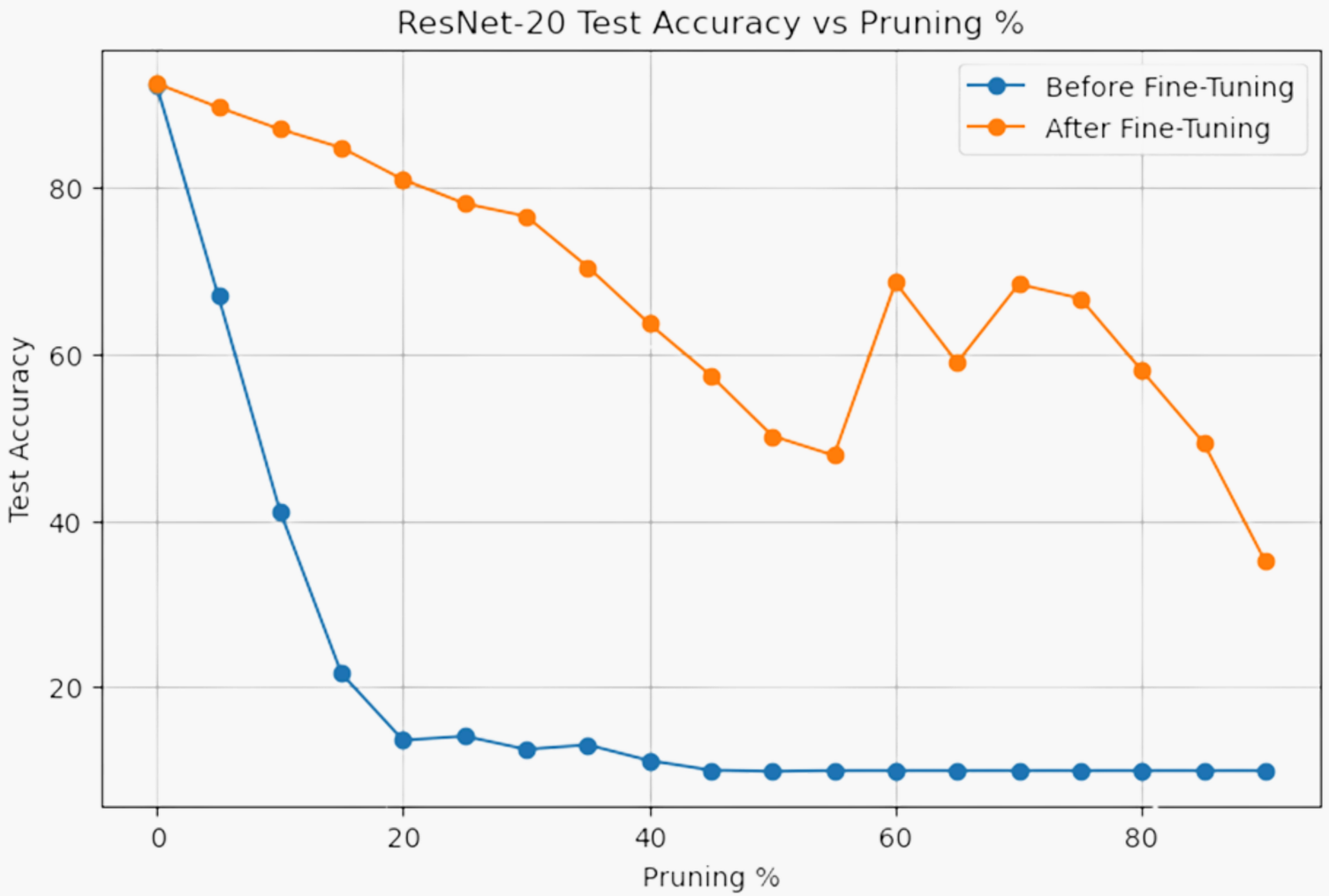
Compressed

2.96 MB, 89.03% accuracy

Compressed Model

93% size reduction with only ~3% drop in accuracy

RESNET-20 COMPRESSION



Pruning severely hurts accuracy

RESNET-20 COMPRESSION

Too small to benefit from quantization

ResNet-20 CIFAR-10 Quantization Results Summary Table

Format	Model Size (MB)	Trainable Parameters (K)	Inference Time (ms/image)	Test Accuracy (%)
FP32 (TRT)	1.40	272.47	0.24	92.12
FP16	1.07	272.47	0.18	92.12
INT8	2.69	272.47	0.15	92.08

RESNET-18 VS. RESNET-20 (TINY IMAGENET)

ResNet-18: 68.89% (Li et al.)

ResNet-20: 55.40% (Yu)

ResNet-20 falls short on complex tasks

RESULTS

Model	Model Size (MB)	Accuracy (%)	Inference Speed (ms/image)
ResNet-18 (FP32)	42.66	92.02	0.61 (TRT)
ResNet-18 Pruned 60% + INT8	2.96	89.03	0.12 (TRT)
ResNet-20 (FP32)	1.40	92.12	0.24 (TRT)

CONCLUSIONS

- Large models (e.g., ResNet-18) compress well with little accuracy loss.

- Small models (e.g., ResNet-20) offer limited compressibility and degrade faster.

- **Choose model size based on task difficulty**

FUTURE WORK

Explore Quantization-Aware Training (QAT)

Use adaptive, per-layer pruning

Benchmark on edge devices (e.g., SmartPhone)

REFERENCES

- Li et al., "Boosting Discriminative Visual Representation Learning with Scenario-Agnostic Mixup, 2022"
- Yu, Hujia. "Deep Convolutional Neural Networks for Tiny ImageNet Classification", CS231n, Stanford, 2017.
- Dadalto, E. ResNet-18 pretrained on CIFAR-100 Hugging Face. https://huggingface.co/edadaltocg/resnet18_cifar100
- Chen, Y. ResNet-20 – PyTorch Models (CIFAR-10). https://github.com/chenyaofo/pytorch-cifar-models
- Nguyen-Phan, H. PyTorch_CIFAR10 (ResNet-18 for CIFAR-10, untrained). https://github.com/huyvnphan/PyTorch_CIFAR10