Discover the philosophy behind deep learning computing

DeePhi Tech

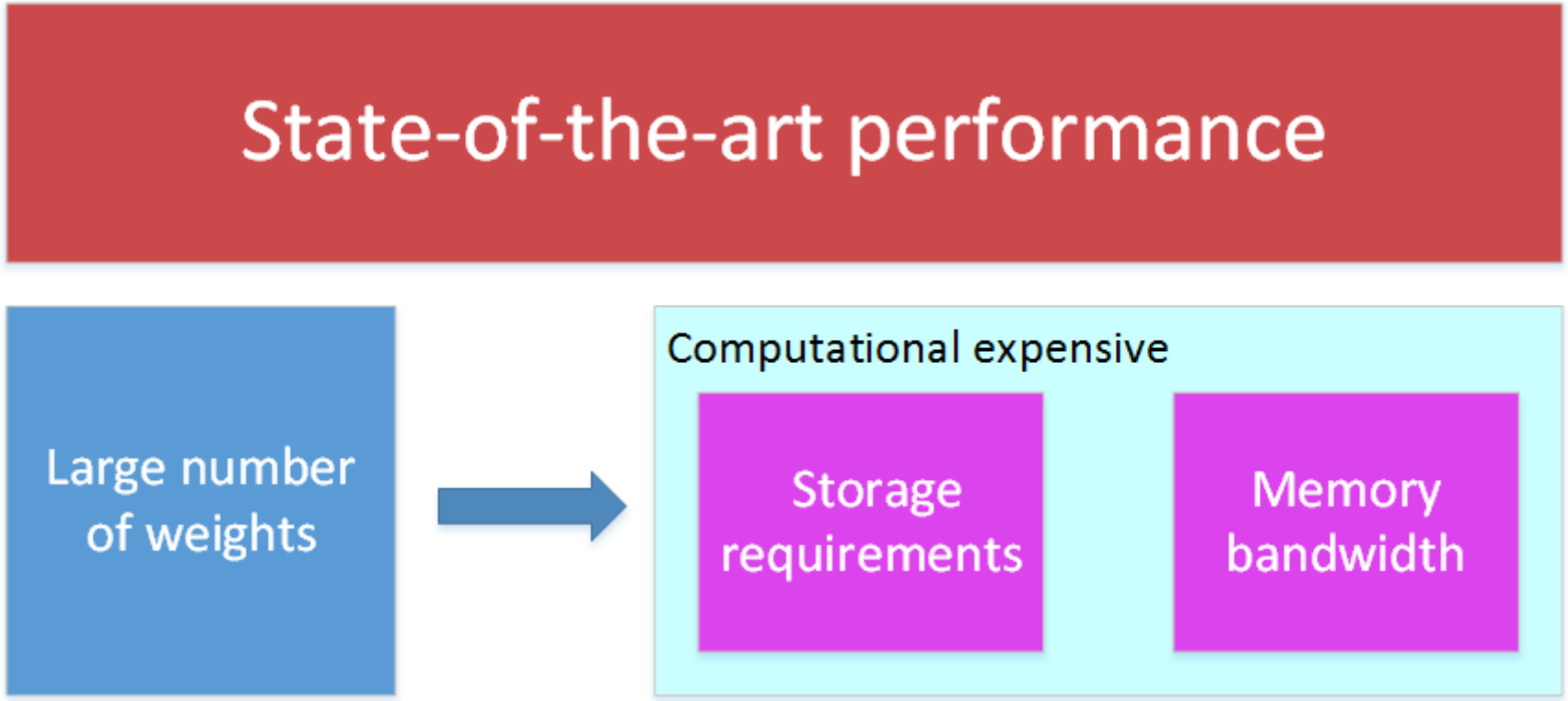
CNN Compression

2017/7/2

Fan Jiang
DeePhi Tech 深鉴科技
fan.jiang@deephi.tech

Background

- Deep Neural Networks



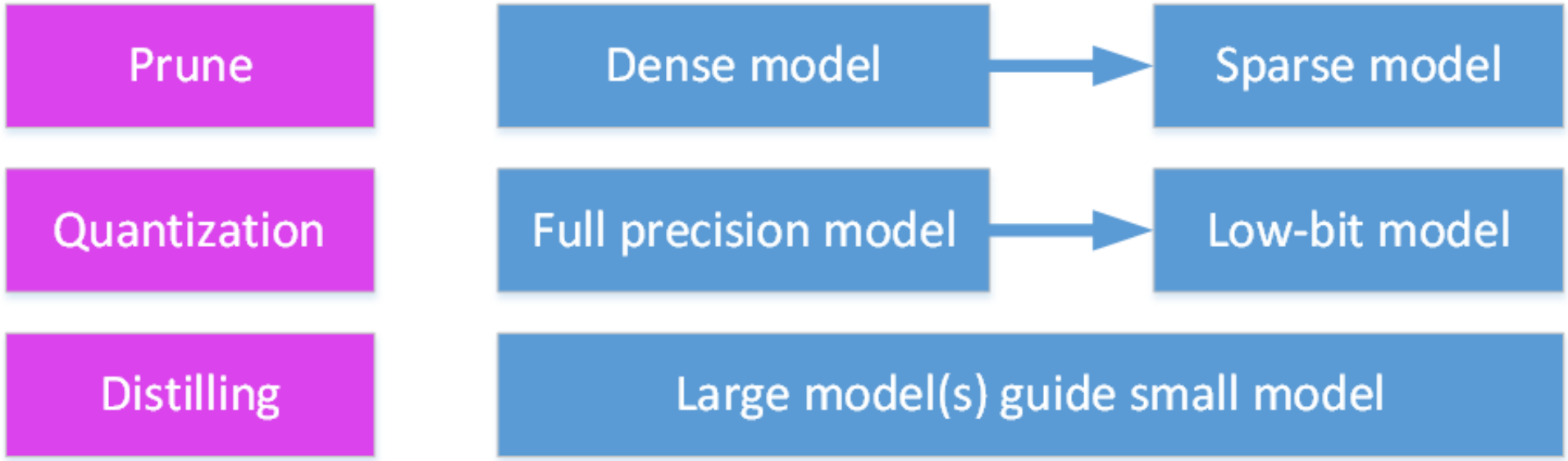
Compression： The Future of Neuron Network Computing

Smaller and Faster	Fewer Computations	Lower Power
Shorter latency in memory read and write	High equivalent performance	Significantly power reduction in memory operation

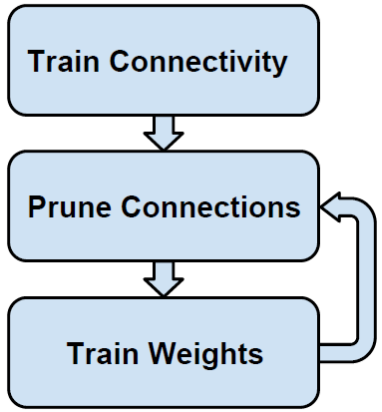
Sparsity will have high priority in future designs.

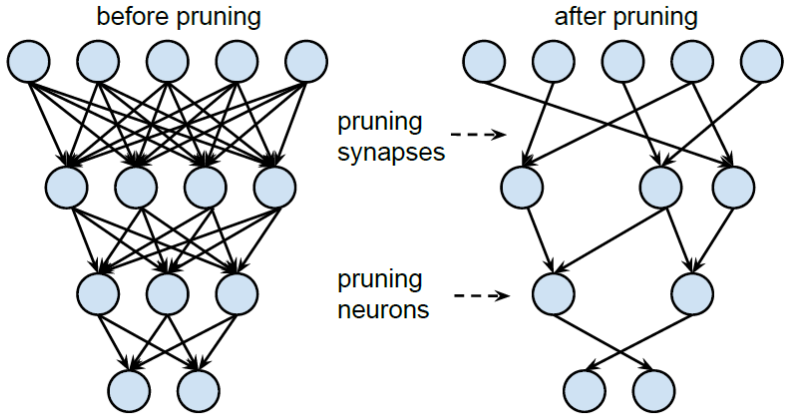
—— Google TPU Paper

Compression techniques

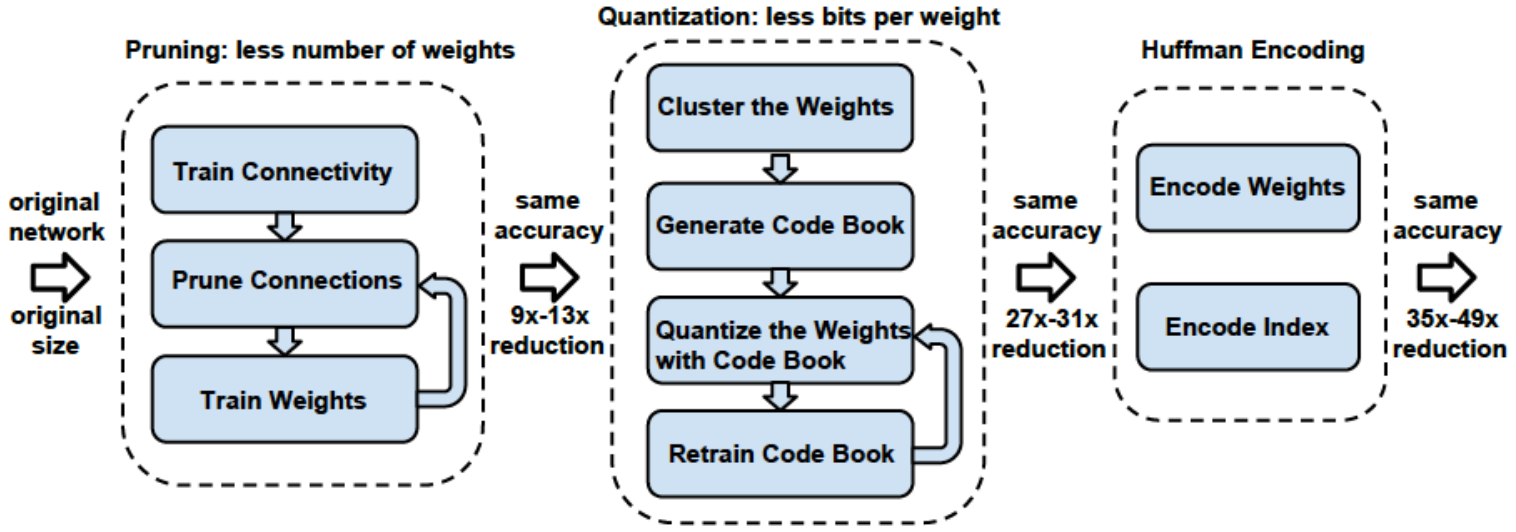


Prune: academic publications (1)



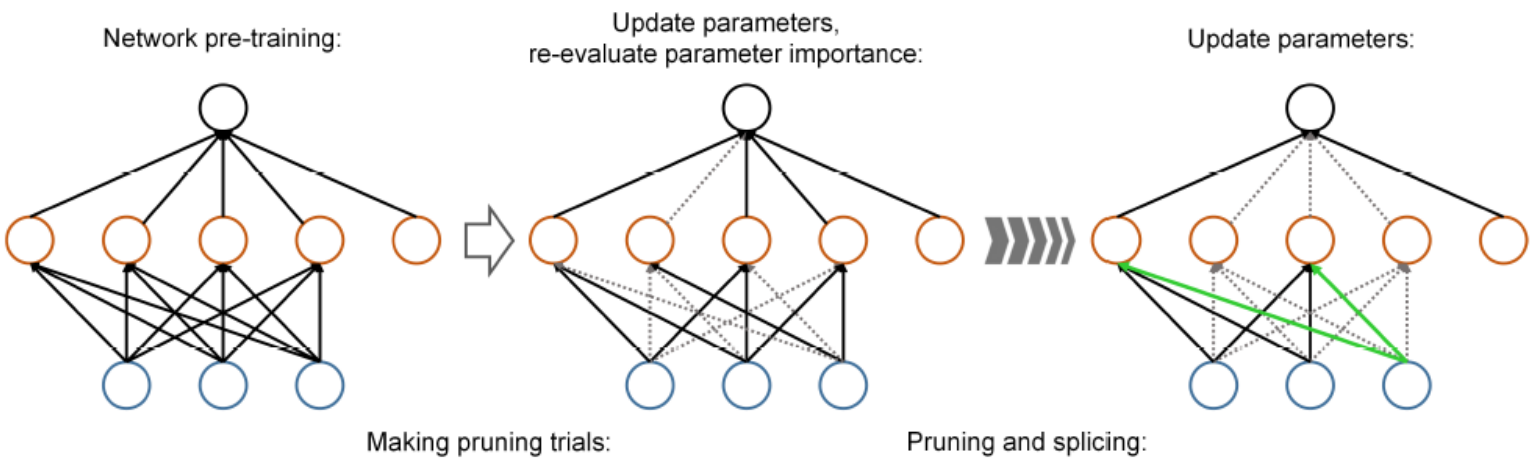


Prune
Song Han NIPS 2015

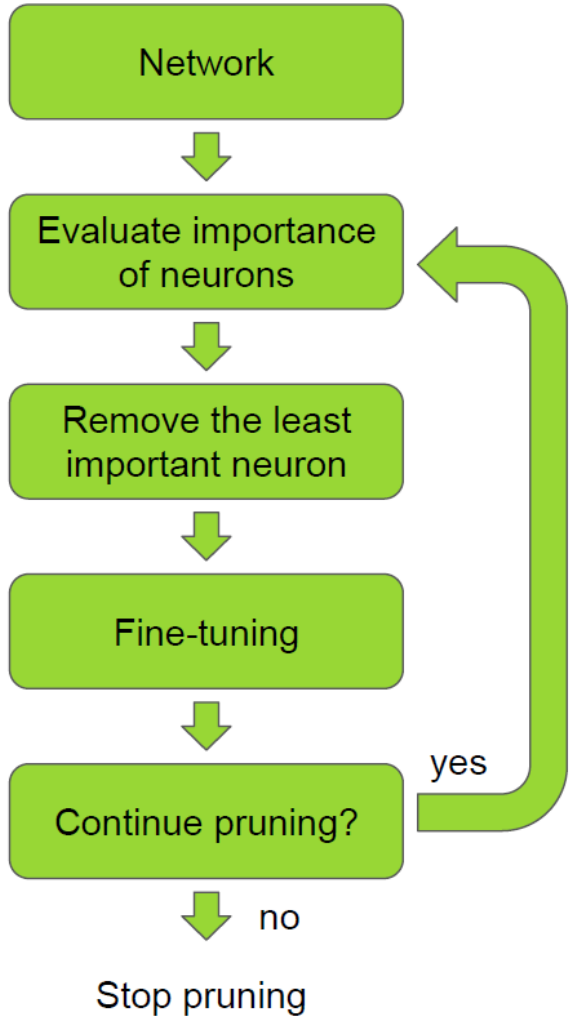


Deep Compression
Song Han ICLR 2016

Prune: academic publications (2)



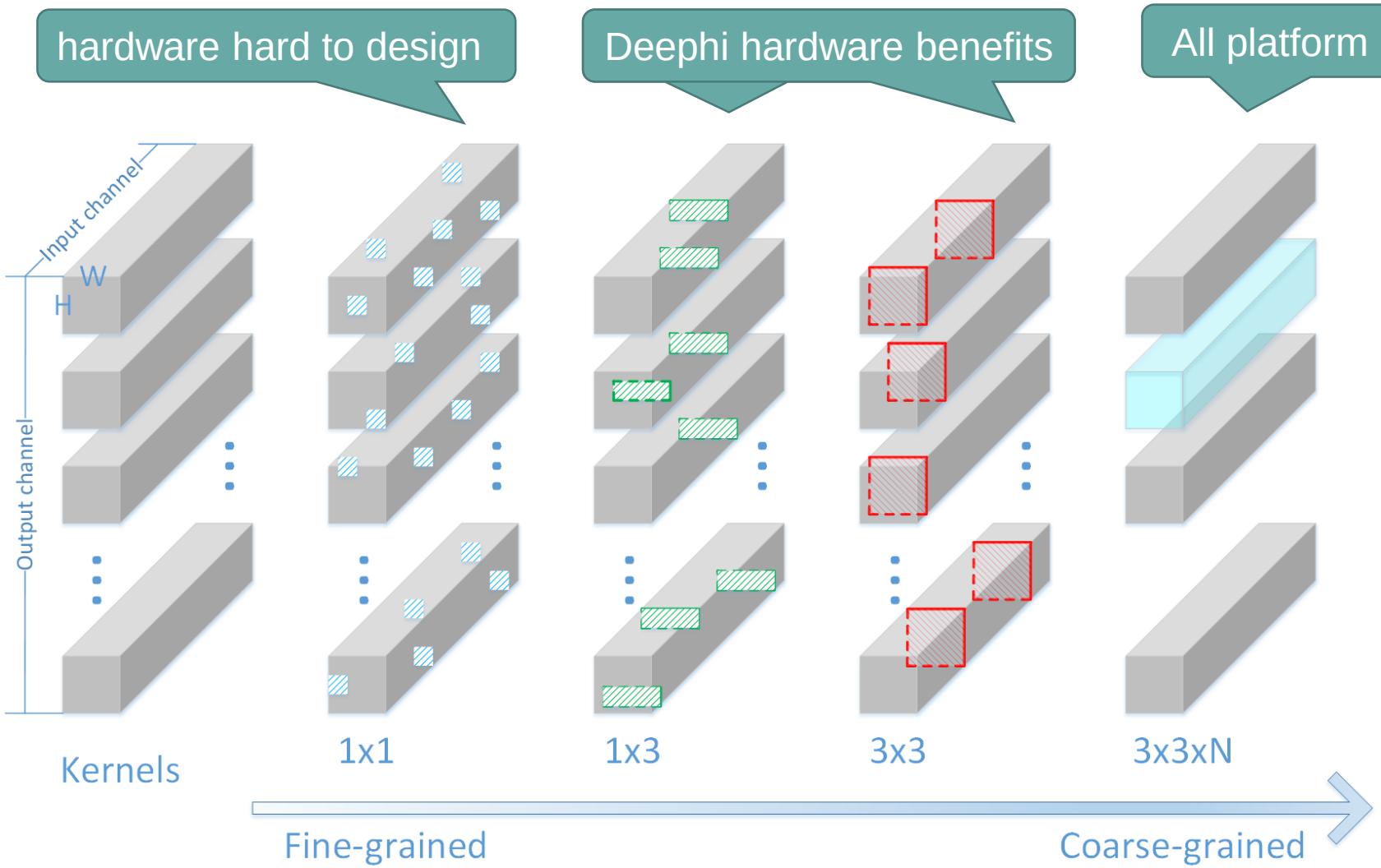
Dynamic Network Surgery, Yiwen Guo, NIPS 2016



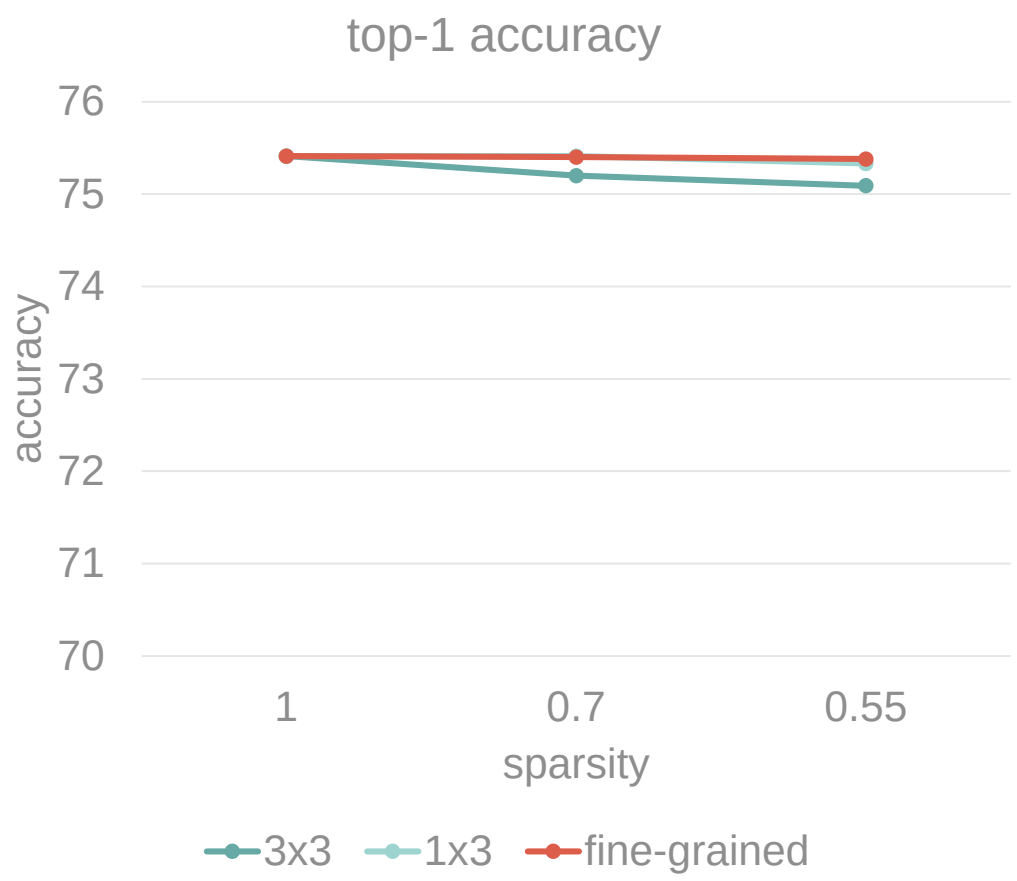
Pruning by nvidia, Pavlo Molchanov, ICLR2017

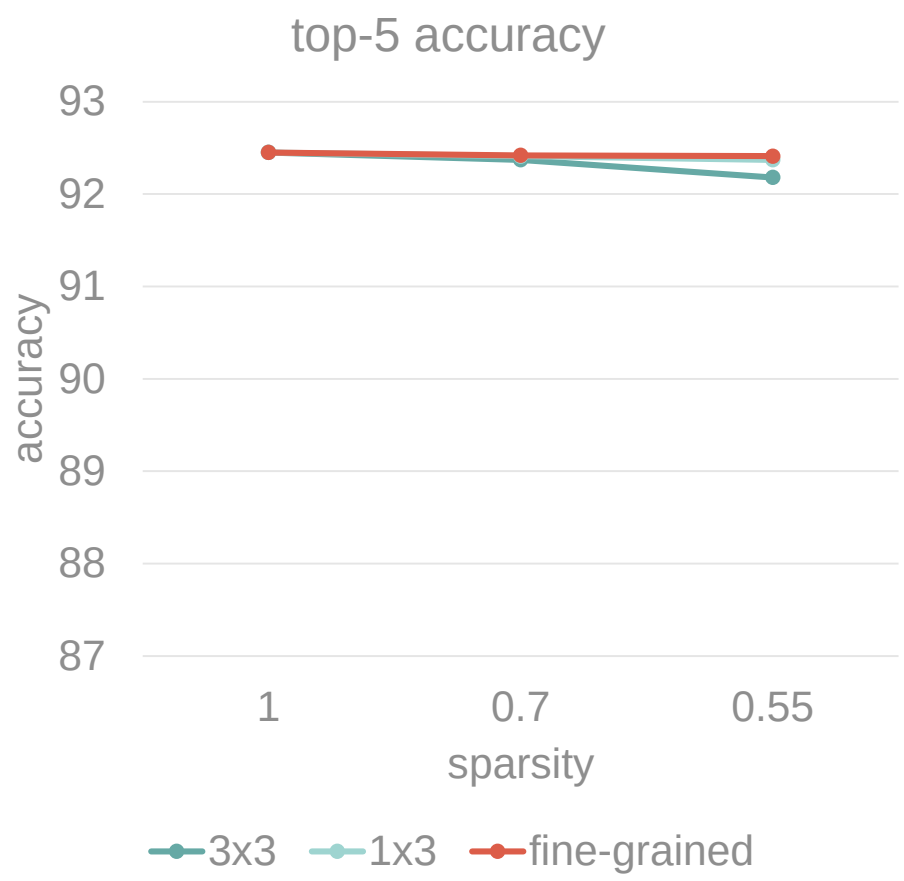
细粒度

Net prune from fine-grained to coarse-grained

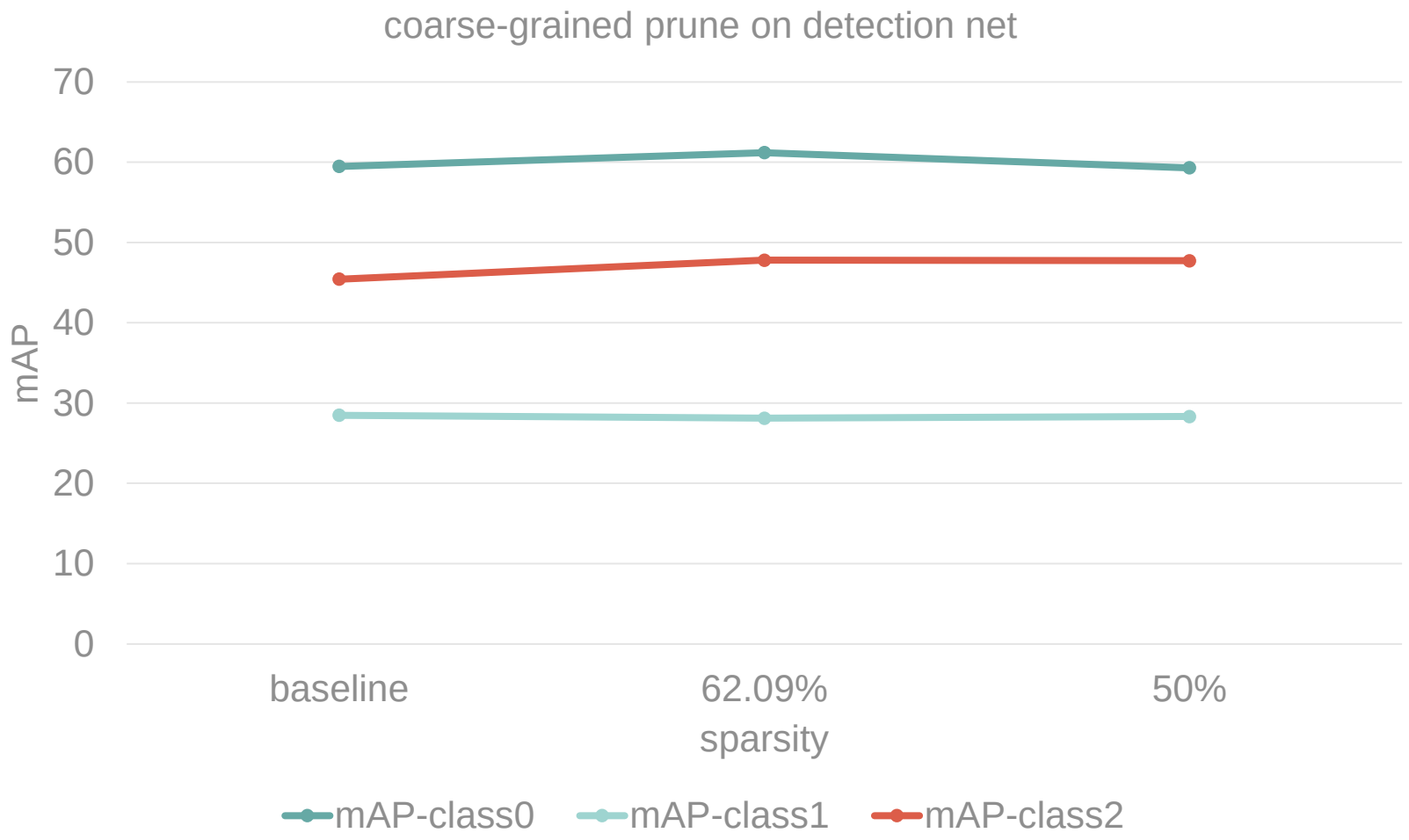


Prune on classification network (resnet-50@imagenet)





Coarse-grained prune (kernel) on detection net



Quantization: from full precision model to low-bit model

- Binaried neural netwroks [Matthieu Courbariaux 2016, arXiv:1602.02830]
- Quantized neural networks [Itay Hubara 2016, arXiv:1609.07061]
- Xnor-net [Mohammad Rastegari 2016, arXiv:1603:05279]
- Ternary weight networks [Fengfu Li 2016, arXiv:1605.04711]
- Trained ternary quantization [Chenzhuo Zhu, ICLR2017, arXiv:1612.01064]
- DoReFa-net [Shuchang Zhou, 2016, arxiv:1606.06160]
- Deep compression [Song Han, ICLR 2016, arXiv:1510.00149]
- Incremental network quantization [Aojun Zhou, ICLR 2017, arXiv:1702.03044]
- Google TPU [ISCA, 2017]
- Nvidia TensorRT

Quantization: weights, activation and gradient

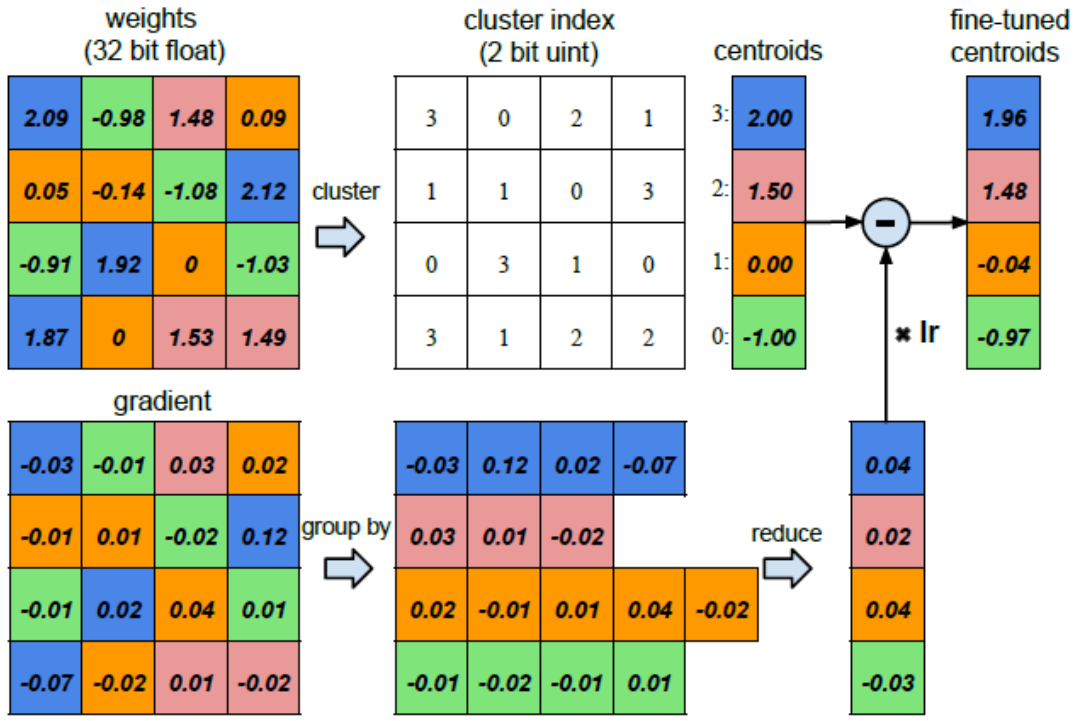
- Balance between performance and accuracy degradation

	Weights (bits)	Activation (bits)	Gradient (bits)
Binaried neural networks	1	1	32
Quantized neural networks	1-8	1-8	6-32
	1	2	6
	4	4	-
Xnor-net	1	32	32
	1	1	32
DoReFa	1	2	4
Ternary weight networks	2	-	32
Trained ternary quantization	2*	32	32
Incremental network quantization	2-5	32/4	32
Google TPU	8	16	-
Deephi DPU	8	8	32

Quantization in Deephi

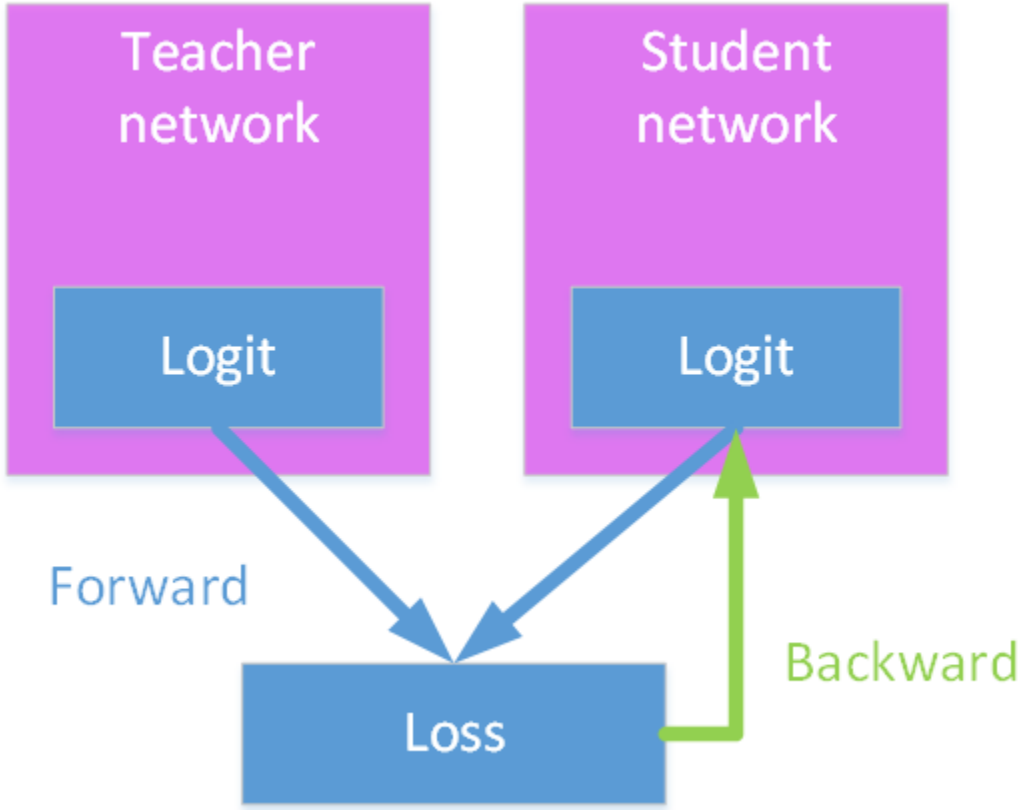
- Uniform quantization
  - 8 bit weights
  - 8 bit activation
  - Full-precision (32bit float) gradient
  - So far, all networks run in DPU use uniform quantization

- Non-uniform quantization
  - K-means cluster
  - Quantize to cluster centroids
  - Software supports non-uniform quantization



Distilling: large model(s) guide small model

- Distilling the Knowledge in a Neural Network [Geoffrey Hinton 2014 NIPS Workshop]
- Transfer knowledge from big model(s) to small model



Distilling together with pruning

- Dataset: cifar10
- Network: Resnet20
- Baseline (dense model) accuracy : 0.9224

Sparsity	Accuracy	Fine-tune	Distillation	Distillation +Fine-tune
0.349	0.7528	0.9186	0.9184	0.9234
0.207	0.3147	0.9098	0.9128	0.9158
0.131	0.1116	0.8742	0.896	0.8810
0.30(coarse)	0.0994	0.8953	0.909	0.9093

Discover the philosophy behind
DEEP LEARNING COMPUTING

THANKS FOR WATCHING