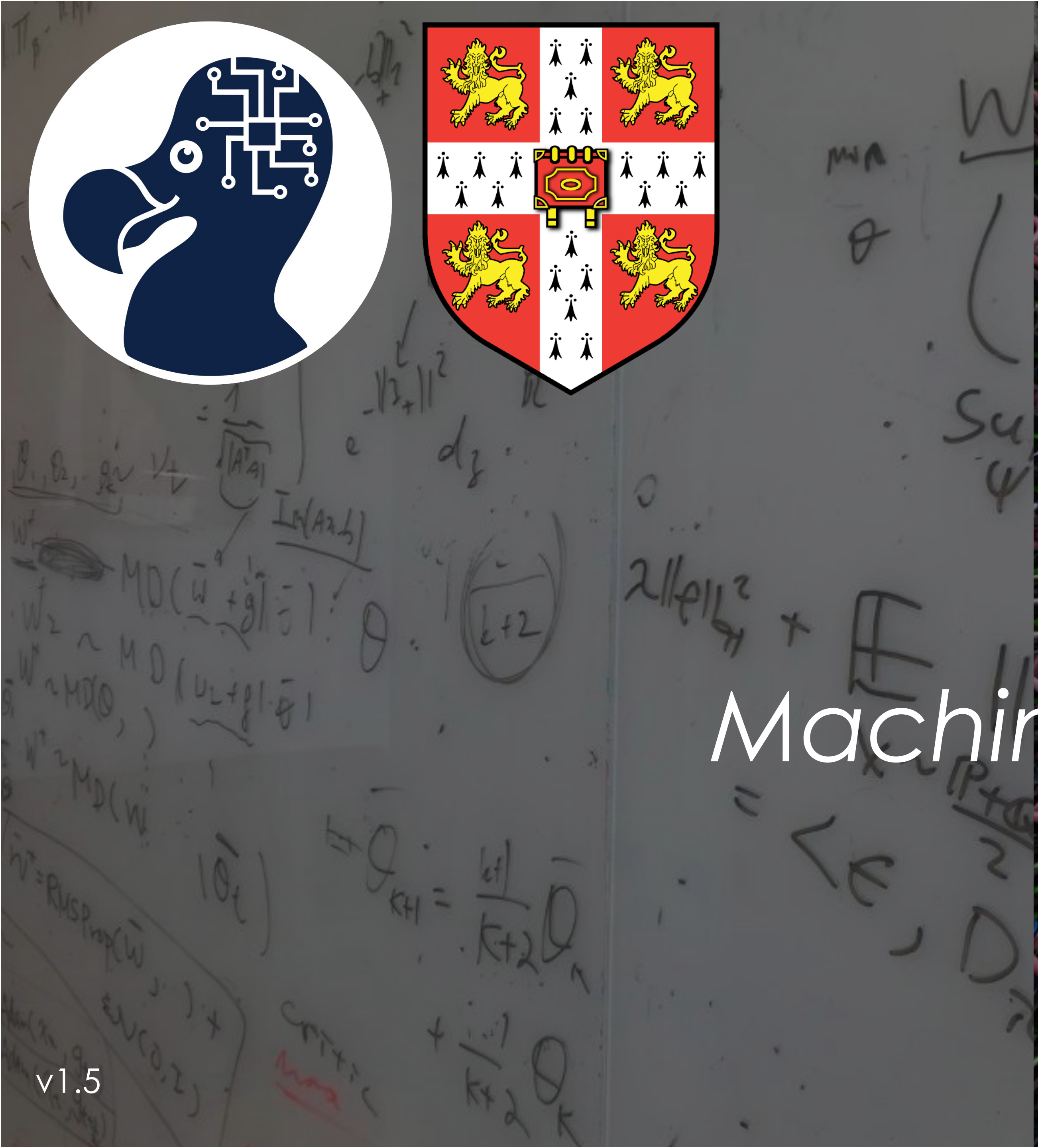
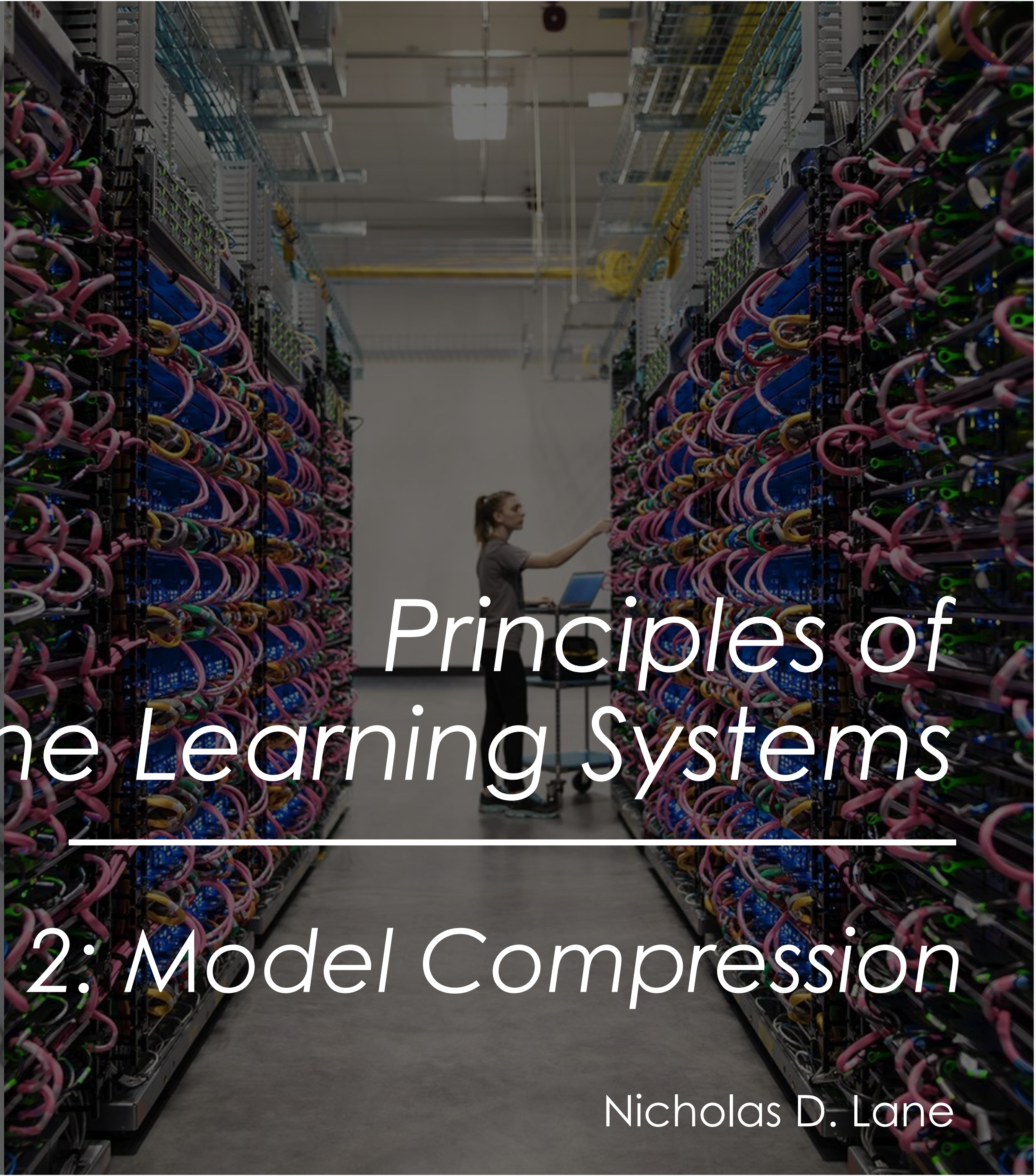




Principles of Machine Learning Systems

2: Model Compression

v1.5

Nicholas D. Lane

Roadmap for Today

Architecture (*Re*-)Design

1. Early Evolution; Decomposition as a tool
2. Efficient Architectures; Example: MobileNet

Architecture Compression

3. Parameter and Channel Pruning
4. Parameter Quantization
5. Knowledge Distillation
6. Compression pipelines

Roadmap for Today

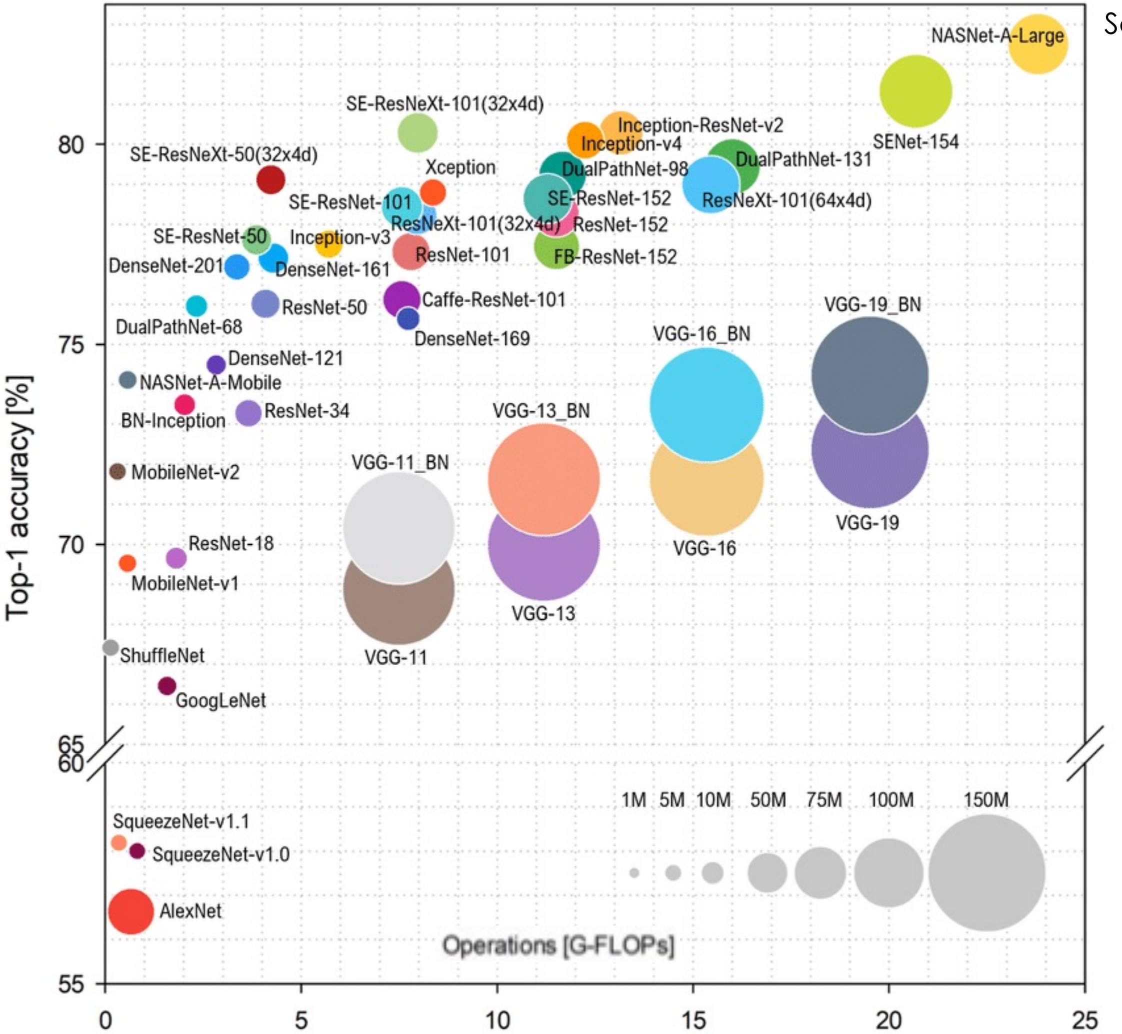
Architecture (*Re*-)Design

1. Early Evolution; Decomposition as a tool
2. Efficient Architectures; Example: MobileNet

Architecture Compression

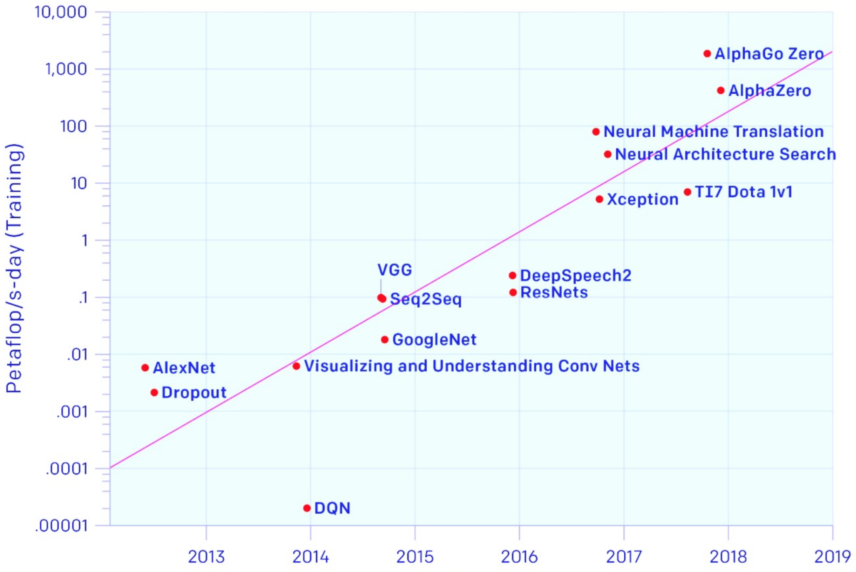
3. Parameter and Channel Pruning
4. Parameter Quantization
5. Knowledge Distillation
6. Compression pipelines

Roadmap for Today

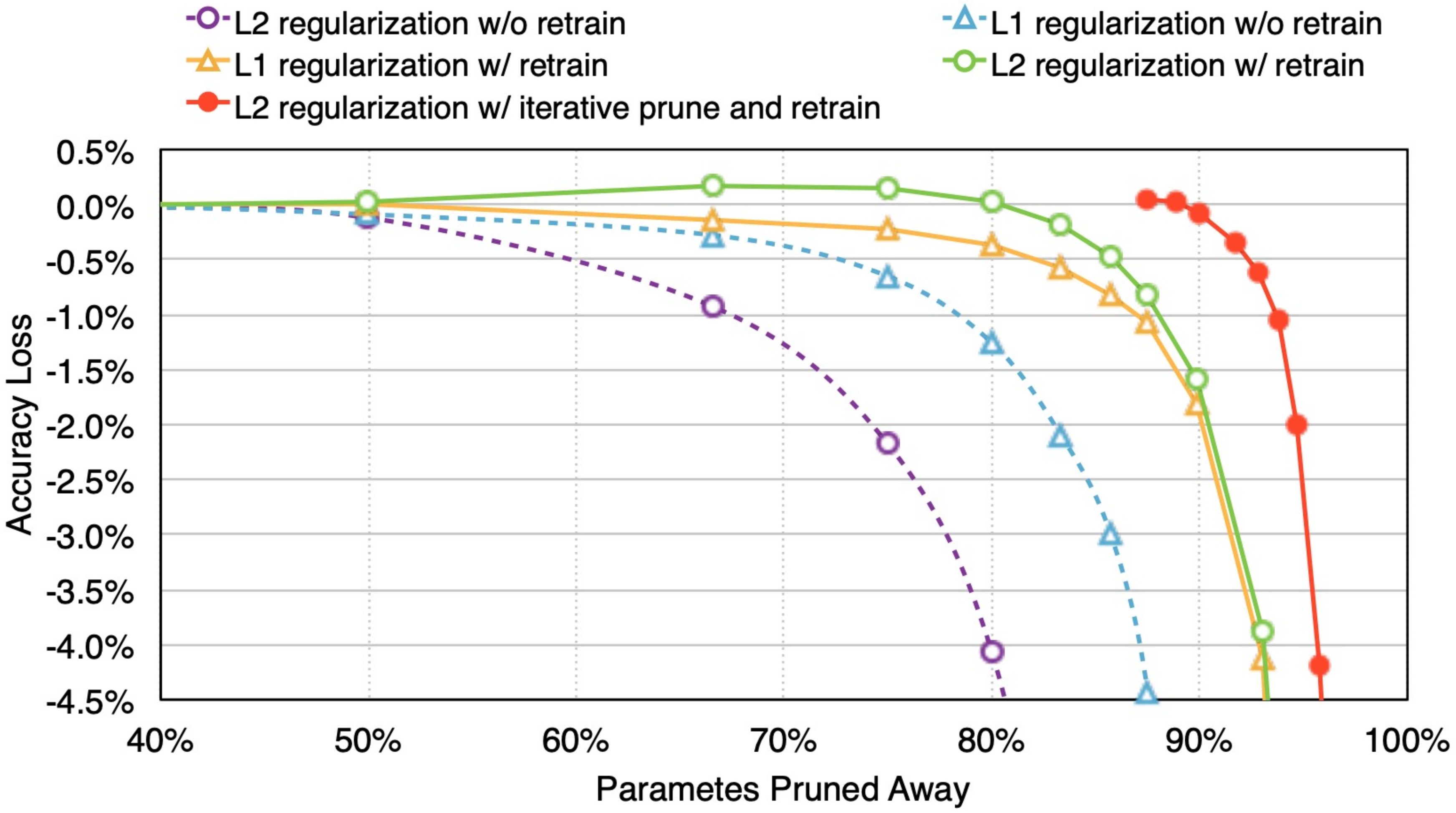


Source: Bianco

Source: OpenAI

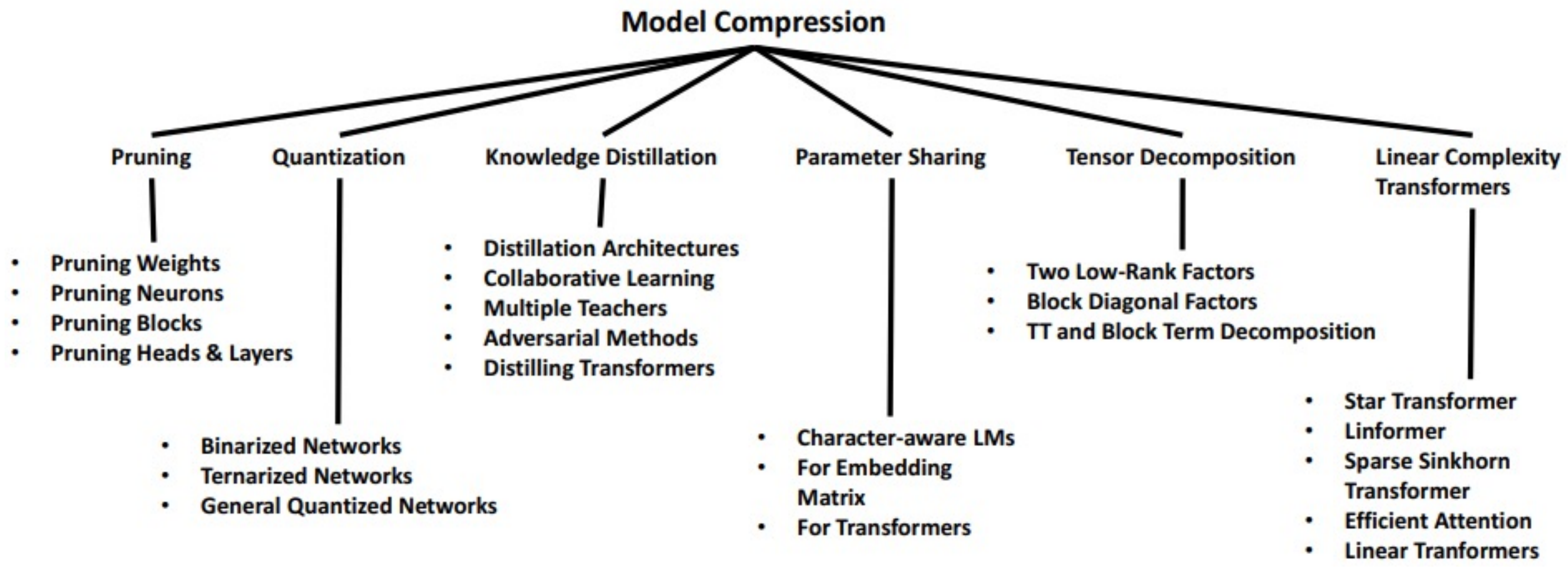


Roadmap for Today



Source: Han

Roadmap for Today



Source: Gupta

Roadmap for Today

Architecture (*Re*-)Design

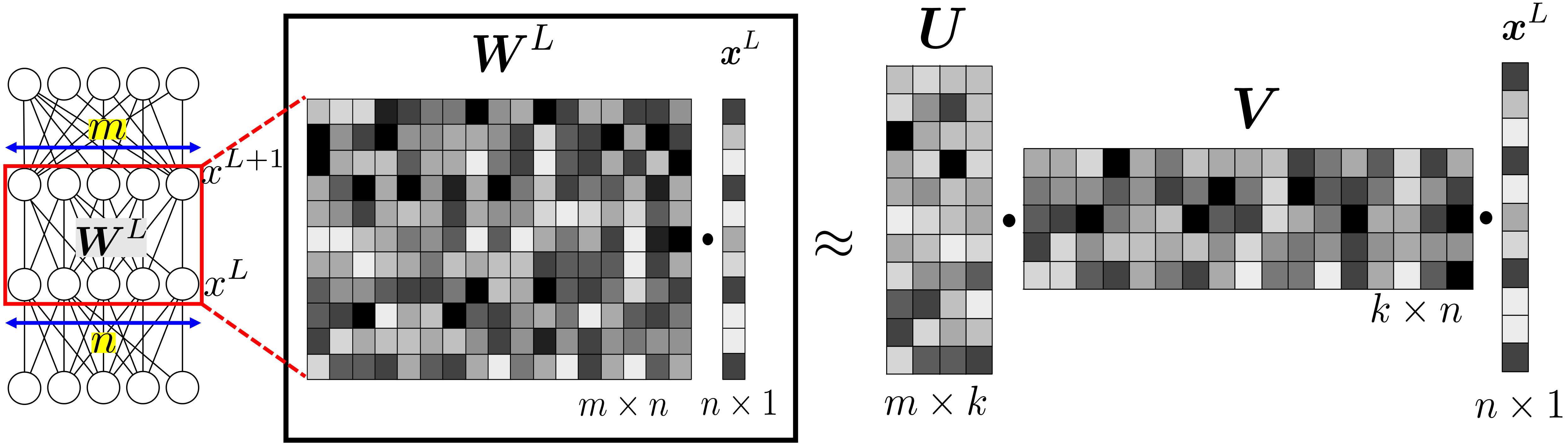
1. **Early Evolution; Decomposition as a tool**
2. Efficient Architectures; Example: MobileNet

Architecture Compression

3. Parameter and Channel Pruning
4. Parameter Quantization
5. Knowledge Distillation
6. Compression pipelines

Matrix Decomposition

Bottleneck: Repeated Matrix Multiplications

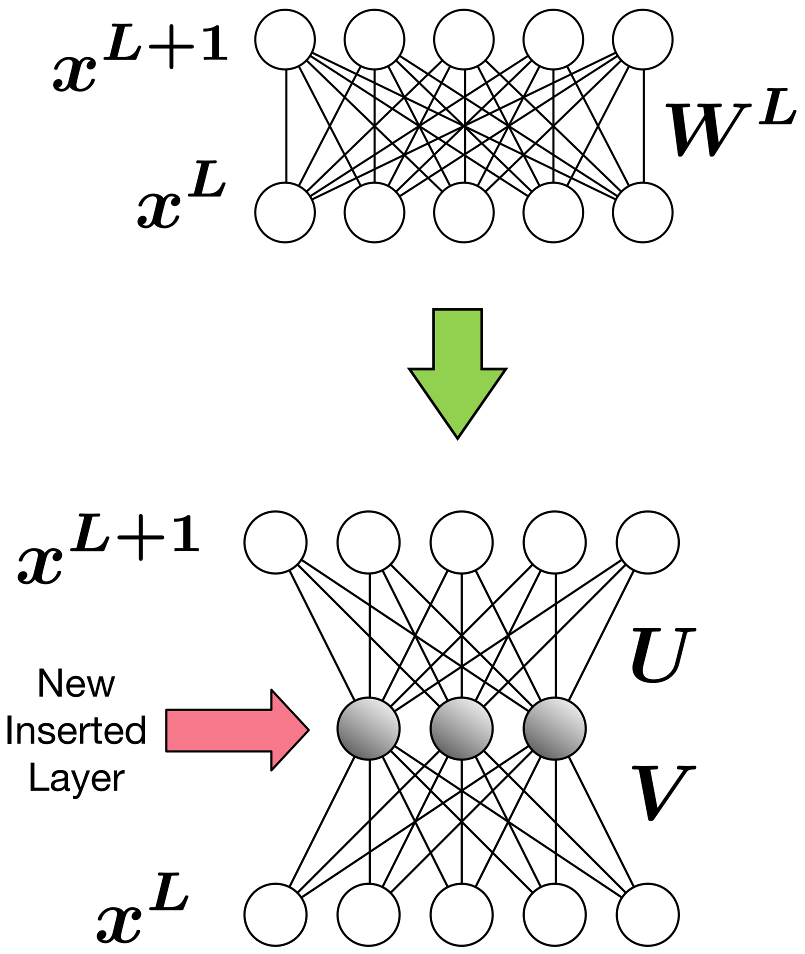


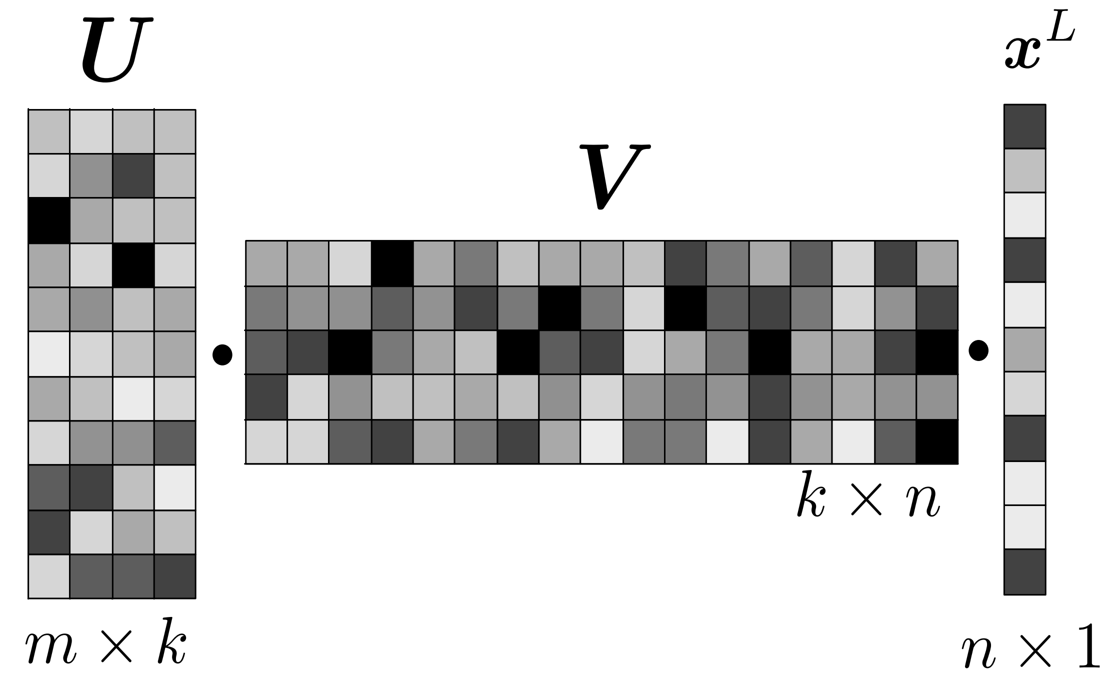
Total Operations: m x n x 1

Total Operations: m x k x 1 + k x n x 1

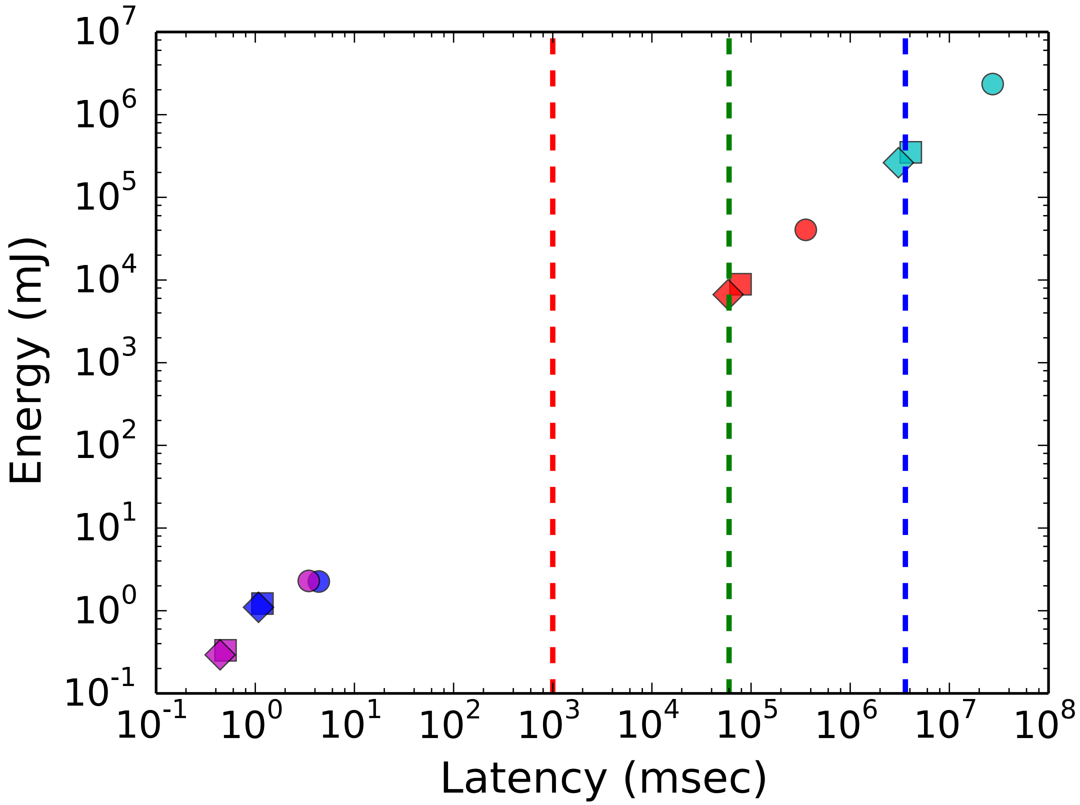
Gain in memory and computations when: $k < \dfrac{m \times n}{m + n}$

Matrix Decomposition



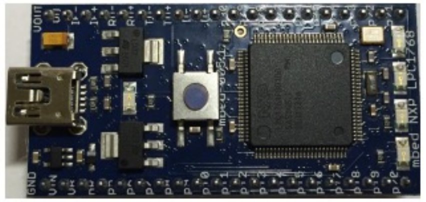


Decomposition Benefits



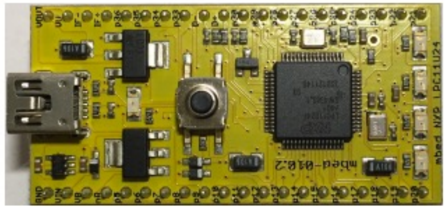
Google SpeakerID Model (FC Layers)

32 KB



ARM Cortex M3

16 KB

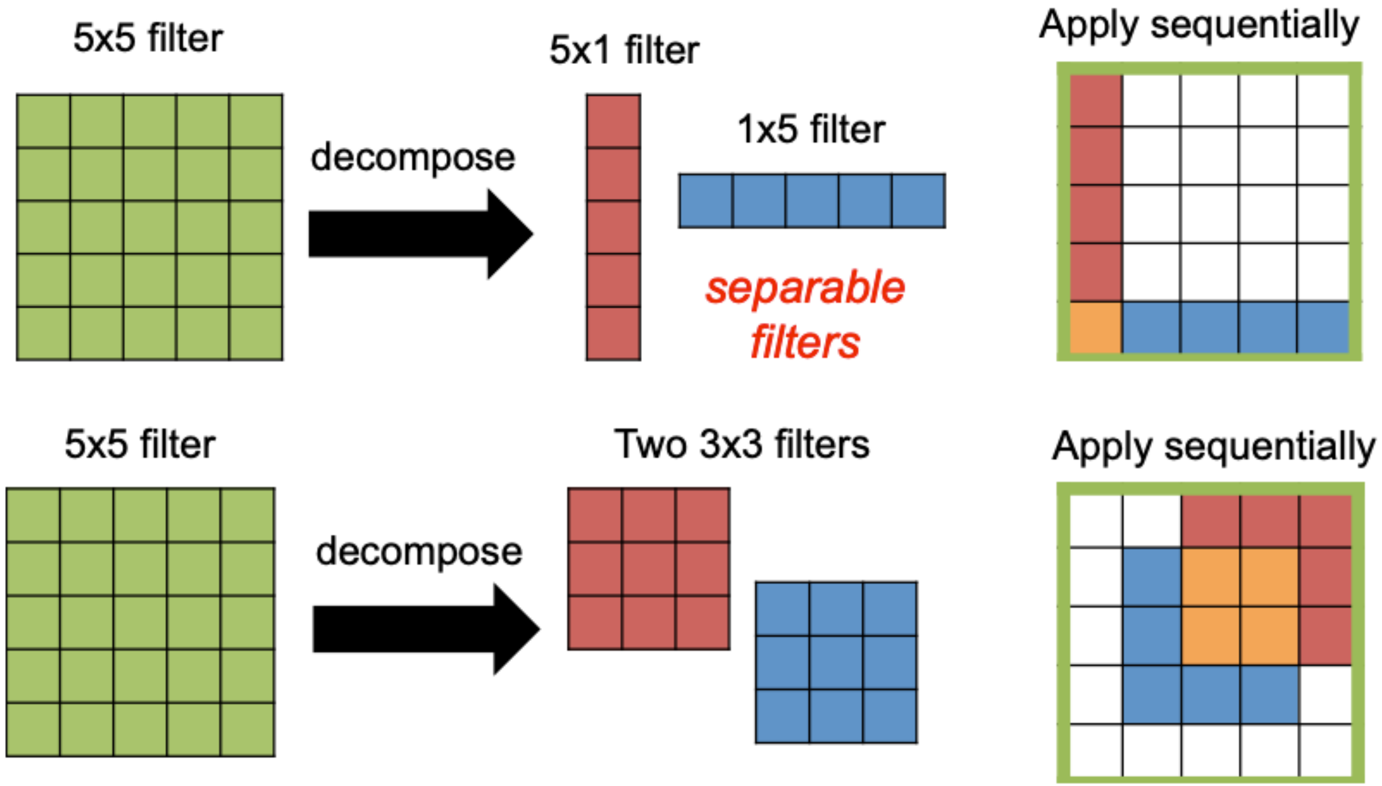


ARM Cortex M0

2-4% degradation in accuracy

CNN Kernel Decomposition

Inception v3
(2014)

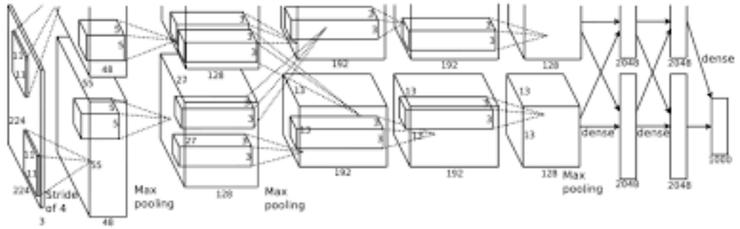


VGG
(2014)

Source: V. Sze

Replacing a large filter with a series of small filters

AlexNet *(2012)*



Roadmap for Today

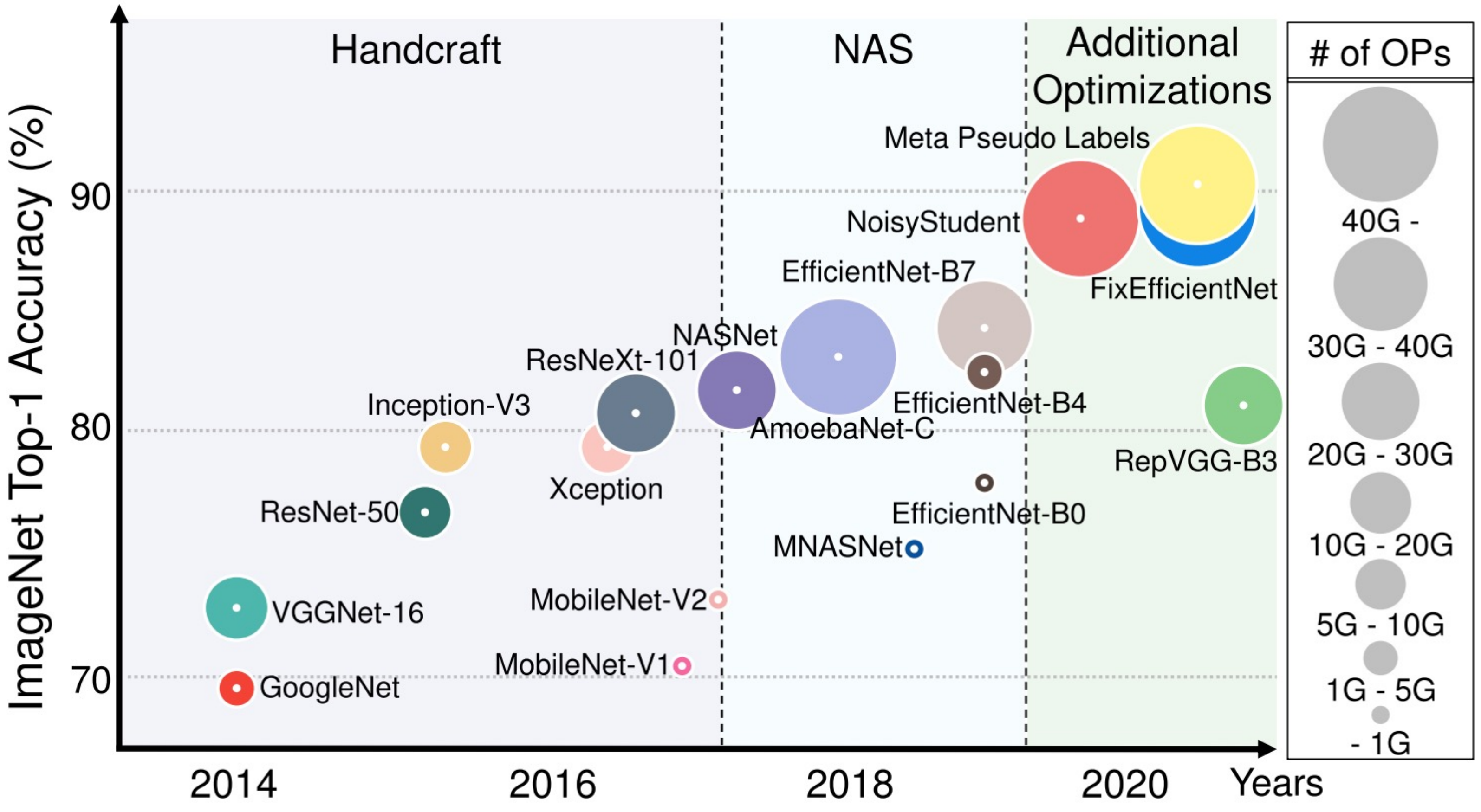
Architecture (*Re*-)Design

1. Early Evolution; Decomposition as a tool
2. **Efficient Architectures; Example: MobileNet**

Architecture Compression

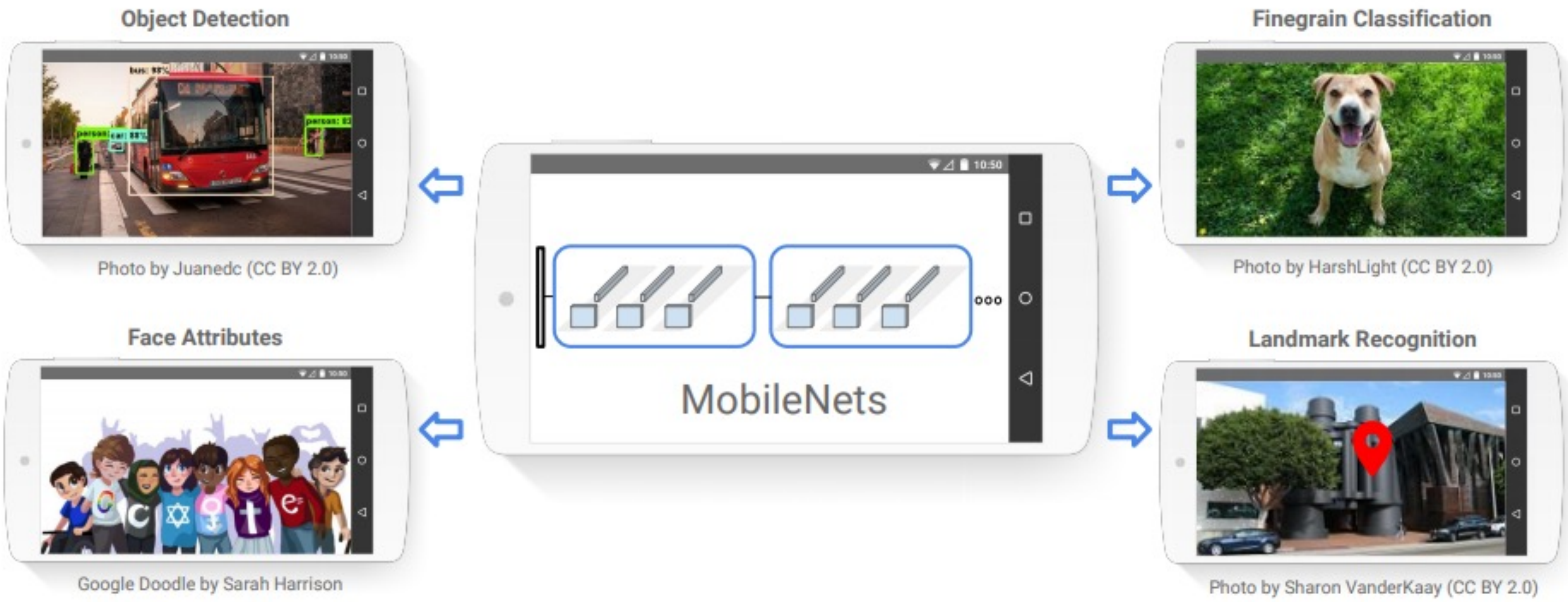
3. Parameter and Channel Pruning
4. Parameter Quantization
5. Knowledge Distillation
6. Compression pipelines

Efficient Architecture Evolution



Source: Lee

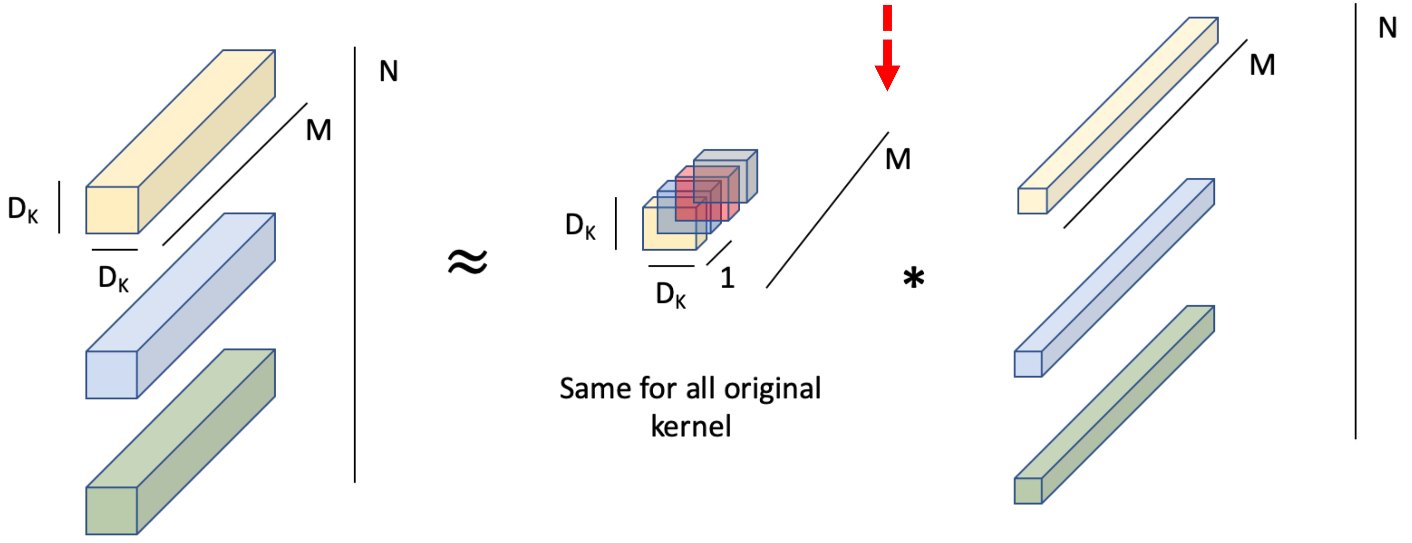
MobileNet Example



Source: Google

MobileNet Block

Type / Stride	Filter Shape	Input Size
Conv / s2	$3 \times 3 \times 3 \times 32$	$224 \times 224 \times 3$
Conv dw / s1	$3 \times 3 \times 32$ dw	$112 \times 112 \times 32$
Conv / s1	$1 \times 1 \times 32 \times 64$	$112 \times 112 \times 32$
Conv dw / s2	$3 \times 3 \times 64$ dw	$112 \times 112 \times 64$
Conv / s1	$1 \times 1 \times 64 \times 128$	$56 \times 56 \times 64$
Conv dw / s1	$3 \times 3 \times 128$ dw	$56 \times 56 \times 128$
Conv / s1	$1 \times 1 \times 128 \times 128$	$56 \times 56 \times 128$
Conv dw / s2	$3 \times 3 \times 128$ dw	$56 \times 56 \times 128$
Conv / s1	$1 \times 1 \times 128 \times 256$	$28 \times 28 \times 128$
Conv dw / s1	$3 \times 3 \times 256$ dw	$28 \times 28 \times 256$
Conv / s1	$1 \times 1 \times 256 \times 256$	$28 \times 28 \times 256$
Conv dw / s2	$3 \times 3 \times 256$ dw	$28 \times 28 \times 256$
Conv / s1	$1 \times 1 \times 256 \times 512$	$14 \times 14 \times 256$
$5\times$ Conv dw / s1	$3 \times 3 \times 512$ dw	$14 \times 14 \times 512$
$5\times$ Conv / s1	$1 \times 1 \times 512 \times 512$	$14 \times 14 \times 512$
Conv dw / s2	$3 \times 3 \times 512$ dw	$14 \times 14 \times 512$
Conv / s1	$1 \times 1 \times 512 \times 1024$	$7 \times 7 \times 512$
Conv dw / s2	$3 \times 3 \times 1024$ dw	$7 \times 7 \times 1024$
Conv / s1	$1 \times 1 \times 1024 \times 1024$	$7 \times 7 \times 1024$
Avg Pool / s1	Pool 7×7	$7 \times 7 \times 1024$
FC / s1	1024×1000	$1 \times 1 \times 1024$
Softmax / s1	Classifier	$1 \times 1 \times 1000$



Original Independent Kernels

One for each kernel

Source: Howard

MobileNet Benefits

Model	ImageNet Accuracy	Million Mult-Adds	Million Parameters
1.0 MobileNet-224	70.6%	569	4.2
GoogleNet	69.8%	1550	6.8
VGG 16	71.5%	15300	138

Model	ImageNet Accuracy	Million Mult-Adds	Million Parameters
Conv MobileNet	71.7%	4866	29.3
MobileNet	70.6%	569	4.2

Source: Howard

Parameter Reduction

$$\frac{\textit{Depthwise Separable}}{\textit{Regular Conv}} = \frac{M*(N + DK^2)}{D_K^2 * M * N} = \frac{1}{D_K^2} + \frac{1}{N}$$

MACs Reduction

$$\frac{\textit{Depthwise Separable}}{\textit{Regular Conv}} = \frac{D_F^2 * M*(N + DK^2)}{D_K^2 * M * DF^2 * N} = \frac{1}{D_K^2} + \frac{1}{N}$$

Roadmap for Today

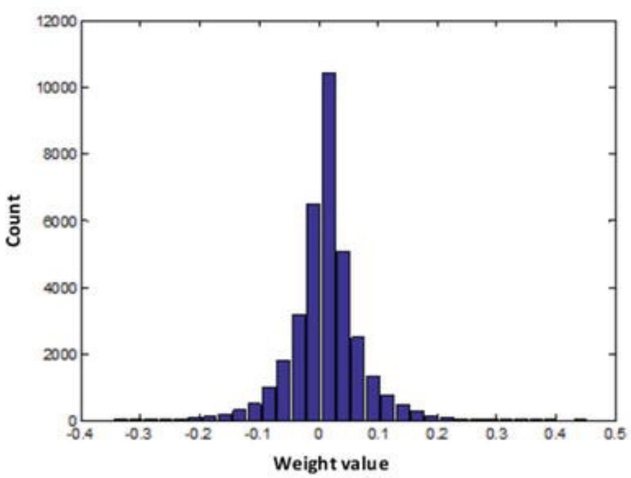
Architecture (*Re*-)Design

1. Early Evolution; Decomposition as a tool
2. Efficient Architectures; Example: MobileNet

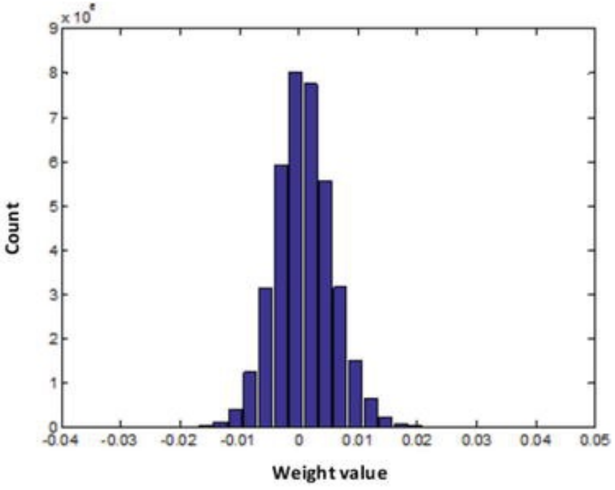
Architecture Compression

3. **Parameter and Channel Pruning**
4. Parameter Quantization
5. Knowledge Distillation
6. Compression pipelines

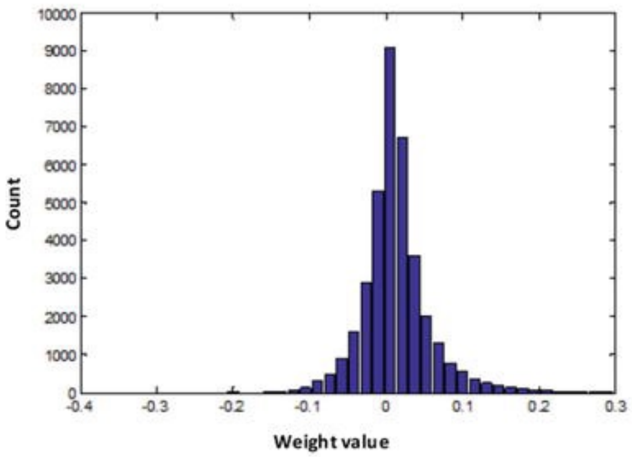
Pruning Justification



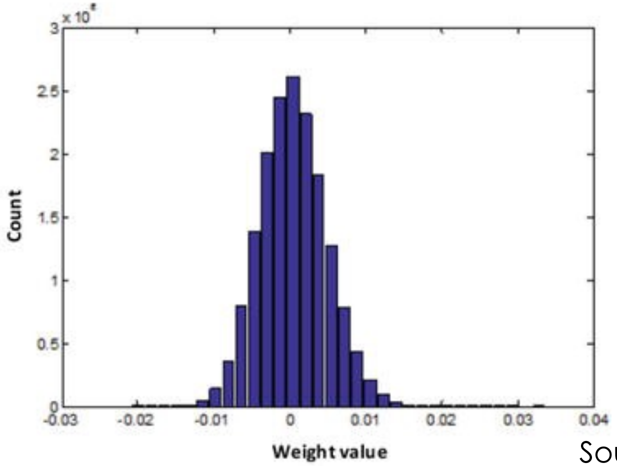
conv1 of AlexNet



fc6 of AlexNet



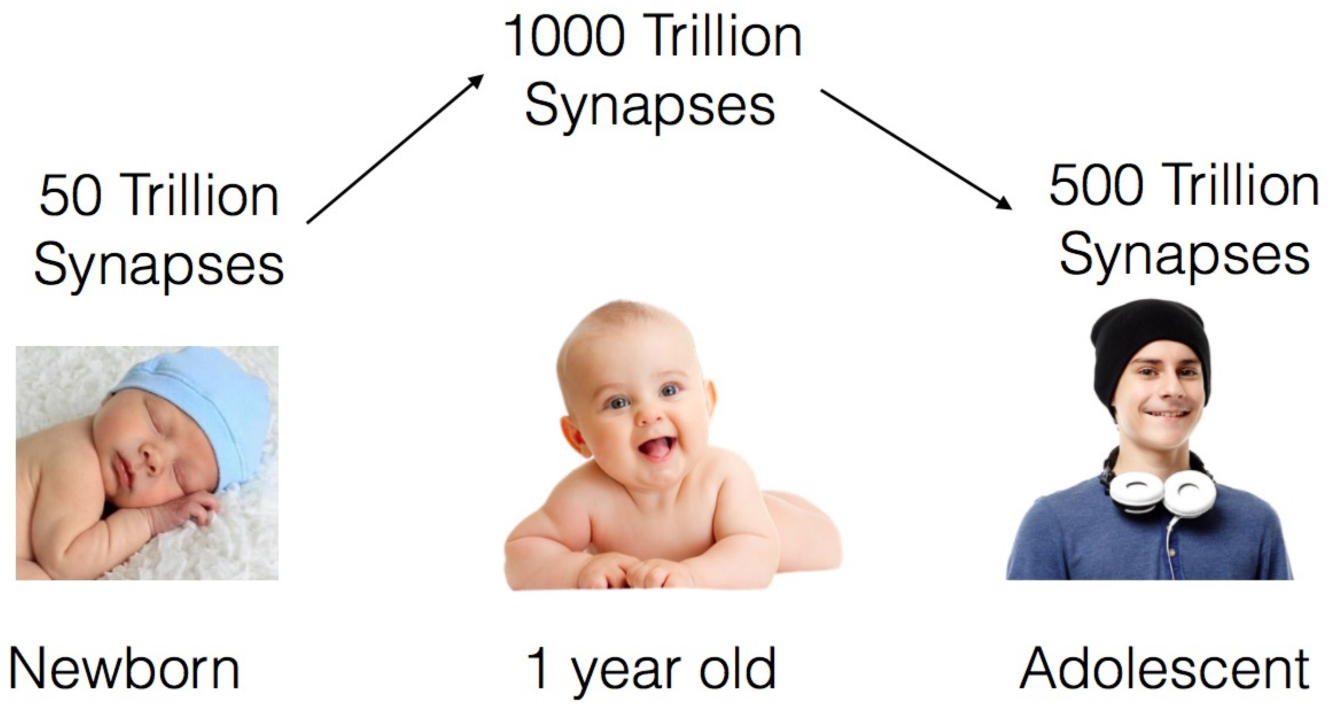
conv12 of VGG-16



fc7 of VGG-16

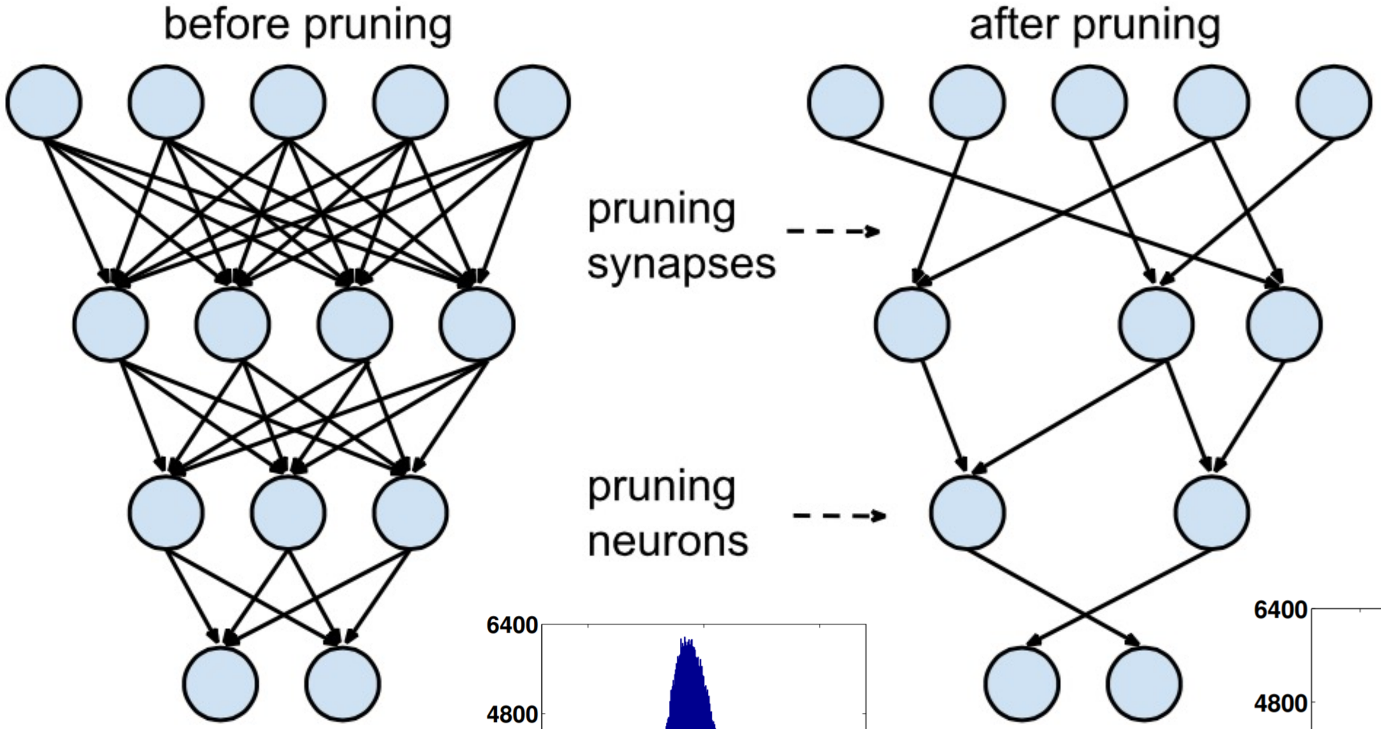
Source: Han

Related Natural Phenomena

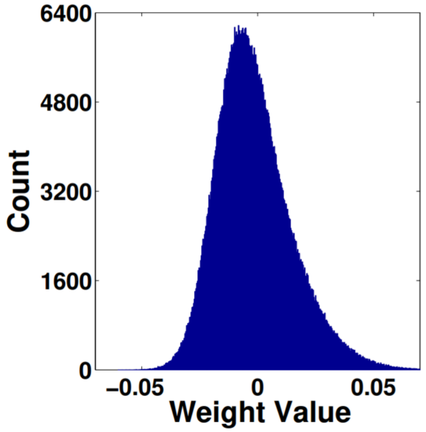


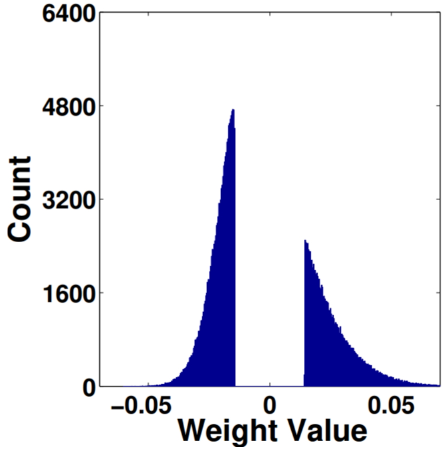
Source: Han

Pruning Algorithm

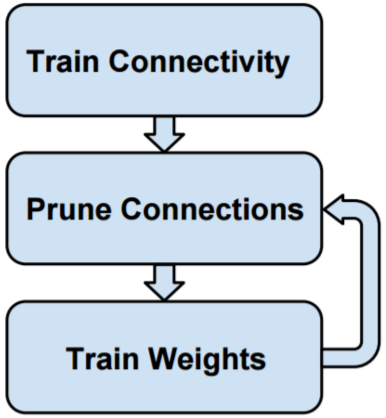


Source: Koul

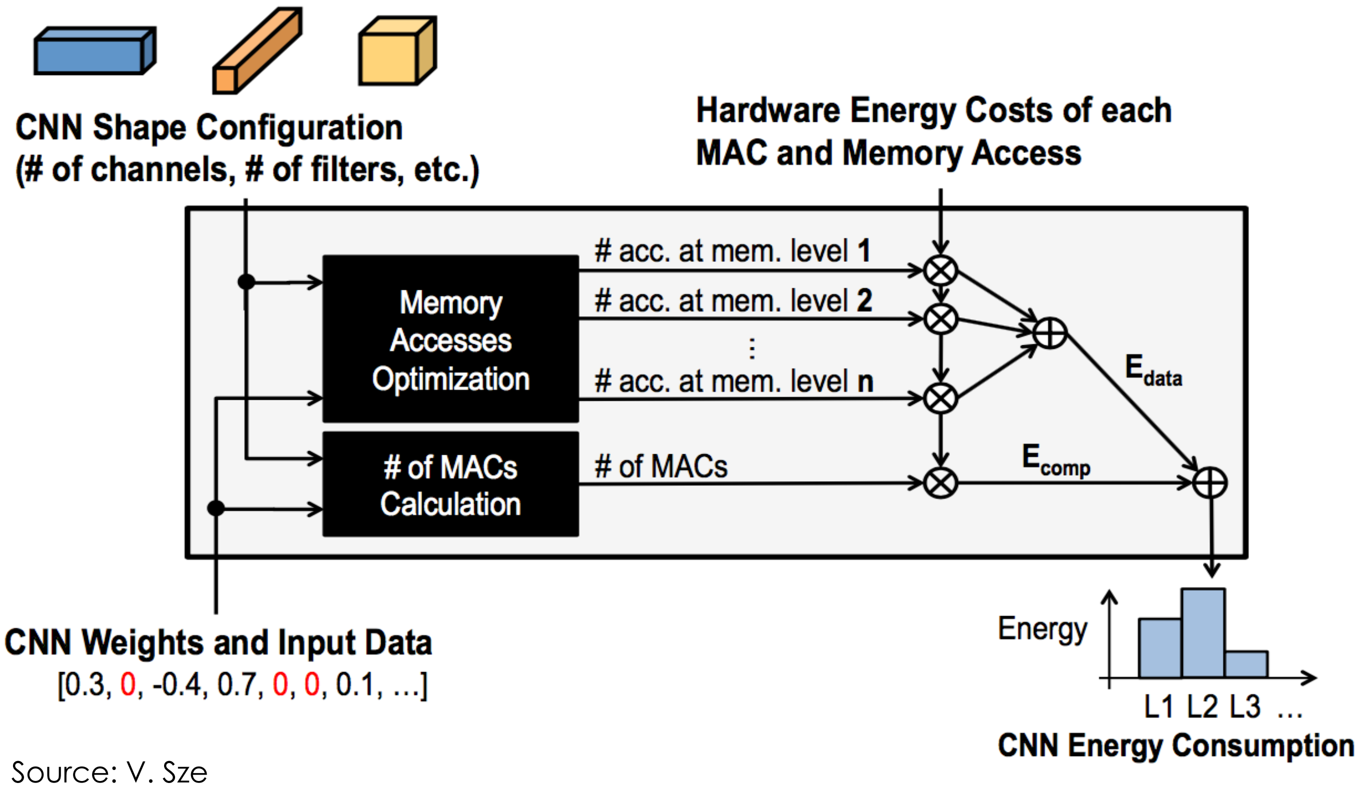




- Range of pruning criteria
- Example: Prune when absolute weight is < a threshold



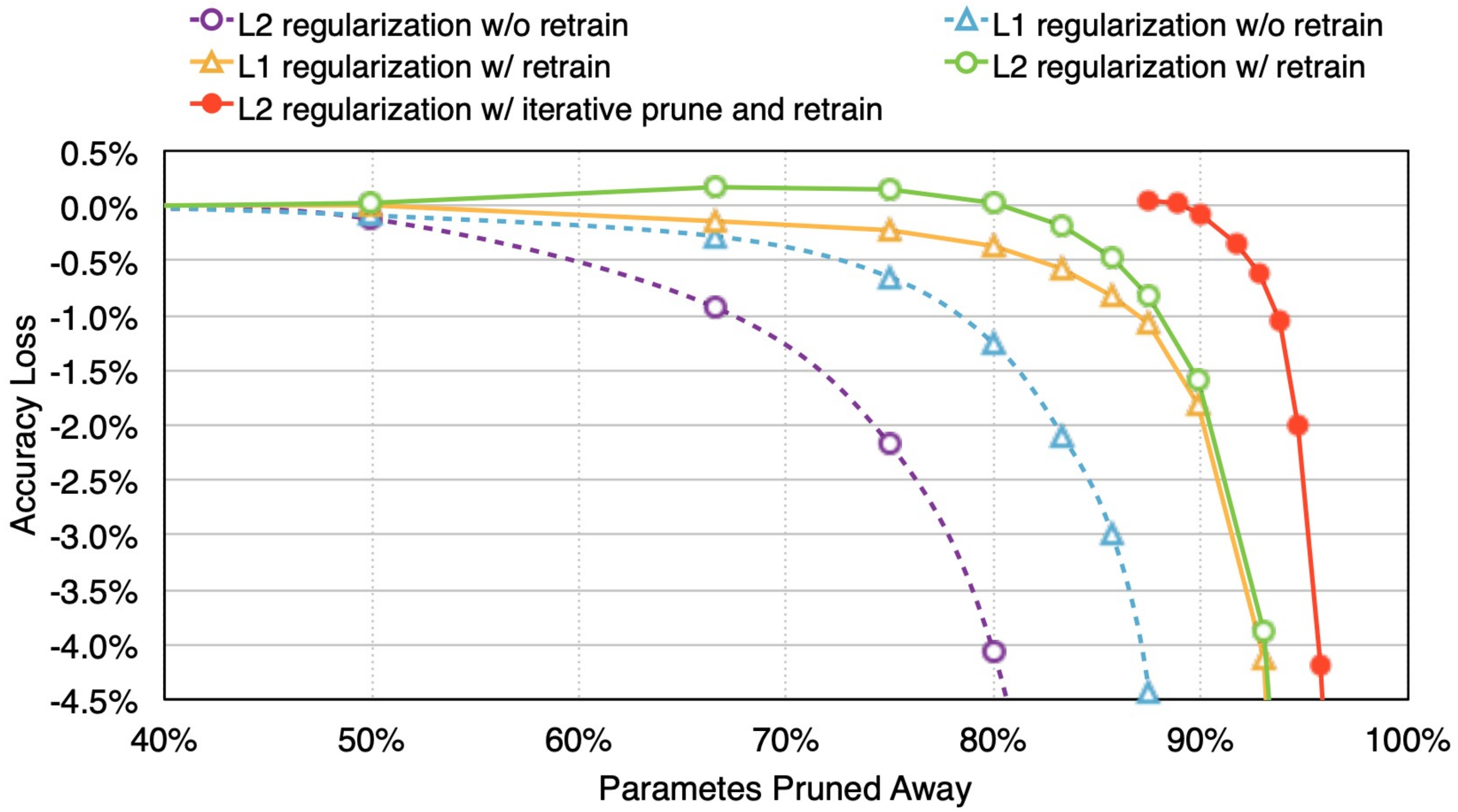
Alternative Pruning Criteria

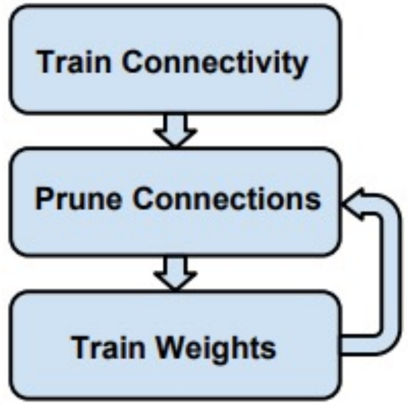


Source: V. Sze

- Magnitude
- Energy
  - Estimation
  - Measurement
- Random
- etc.

Pruning Schedule



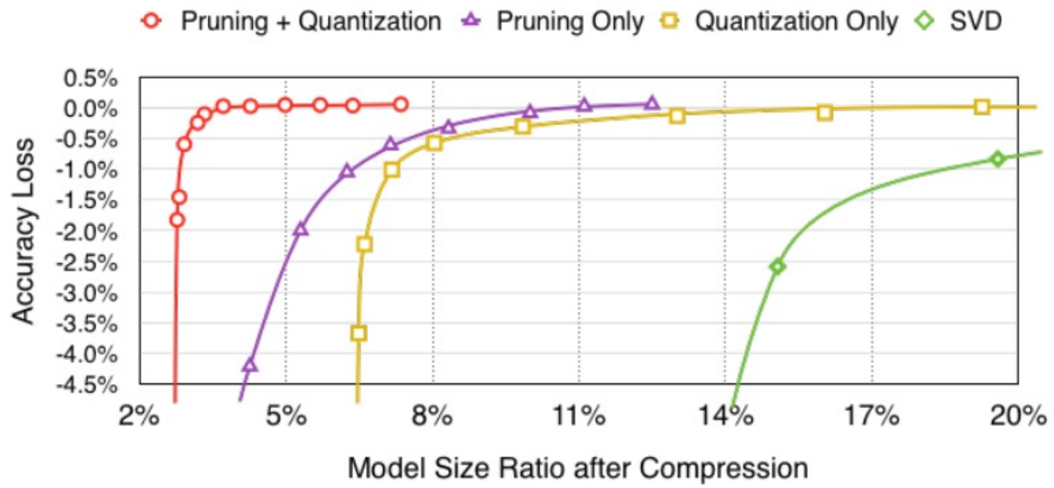


Source: Han

Pruning Performance

Dataset	Model	Unpruned	Prune Ratio	Fine-tuned	Scratch-E	Scratch-B
CIFAR-10	VGG-19	93.50 (±0.11)	30%	93.51 (±0.05)	**93.71** (±0.09)	93.31 (±0.26)
			80%	93.52 (±0.10)	**93.71** (±0.08)	93.64 (±0.09)
			95%	93.34 (±0.13)	93.21 (±0.17)	**93.63** (±0.18)
	PreResNet-110	95.04 (±0.15)	30%	95.06 (±0.05)	94.84 (±0.07)	**95.11** (±0.09)
			80%	**94.55** (±0.11)	93.76 (±0.10)	94.52 (±0.13)
			95%	**92.35** (±0.20)	91.23 (±0.11)	91.55 (±0.34)
	DenseNet-BC-100	95.24 (±0.17)	30%	95.21 (±0.17)	95.22 (±0.18)	**95.23** (±0.14)
			80%	95.04 (±0.15)	94.42 (±0.12)	**95.12** (±0.04)
			95%	**94.19** (±0.15)	92.91 (±0.22)	93.44 (±0.19)
CIFAR-100	VGG-19	71.70 (±0.31)	30%	71.96 (±0.36)	72.81 (±0.31)	**73.30** (±0.25)
			50%	71.85 (±0.30)	73.12 (±0.36)	**73.77** (±0.23)
			95%	70.22 (±0.38)	70.88 (±0.35)	**72.08** (±0.15)
	PreResNet-110	76.96 (±0.34)	30%	76.88 (±0.31)	76.36 (±0.26)	**76.96** (±0.31)
			50%	**76.60** (±0.36)	75.45 (±0.23)	76.42 (±0.39)
			95%	68.55 (±0.51)	68.13 (±0.64)	**68.99** (±0.32)
	DenseNet-BC-100	77.59 (±0.19)	30%	77.23 (±0.05)	77.58 (±0.25)	**77.97** (±0.31)
			50%	77.41 (±0.14)	77.65 (±0.09)	**77.80** (±0.23)
			95%	**73.67** (±0.03)	71.47 (±0.46)	72.57 (±0.37)
ImageNet	VGG-16	73.37	30%	73.68	72.75	**74.02**
			60%	**73.63**	71.50	73.42
	ResNet-50	76.15	30%	**76.06**	74.77	75.70
			60%	**76.09**	73.69	74.91

Source: Wang



Roadmap for Today

Architecture (*Re*-)Design

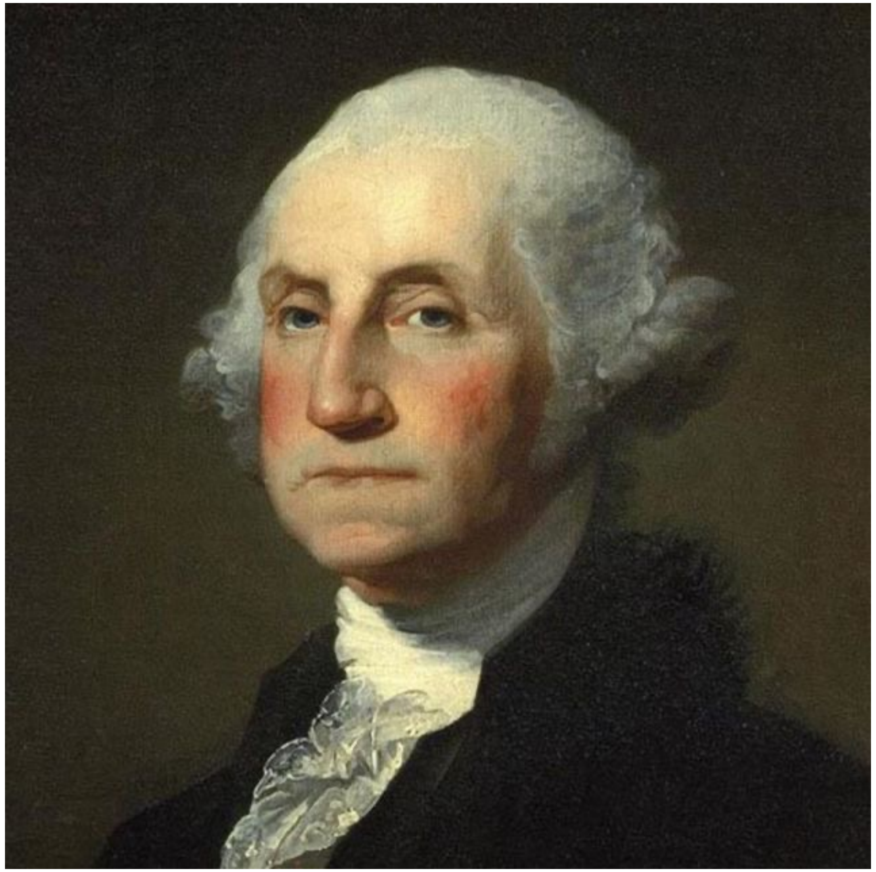
1. Early Evolution; Decomposition as a tool
2. Efficient Architectures; Example: MobileNet

Architecture Compression

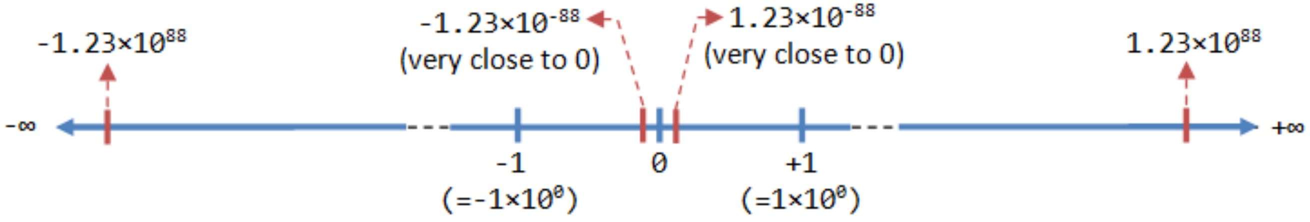
3. Parameter and Channel Pruning
4. **Parameter Quantization**
5. Knowledge Distillation
6. Compression pipelines

Guess who

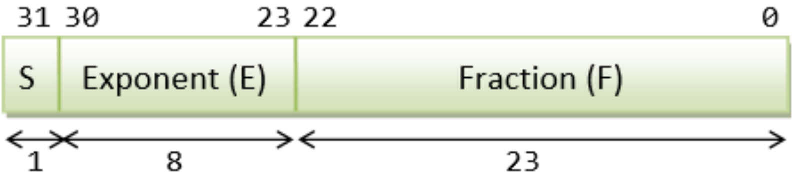
Guess who



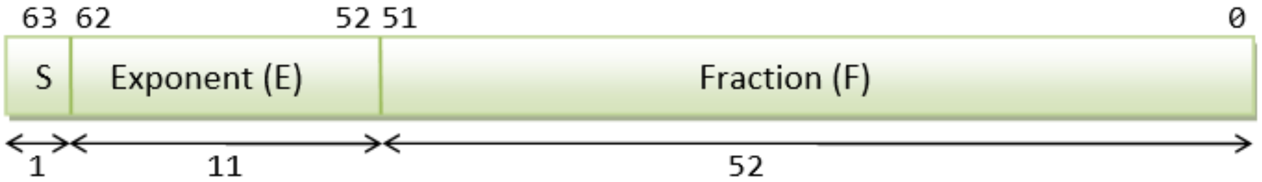
Numerical Representation



Floating-point Numbers (Decimal)

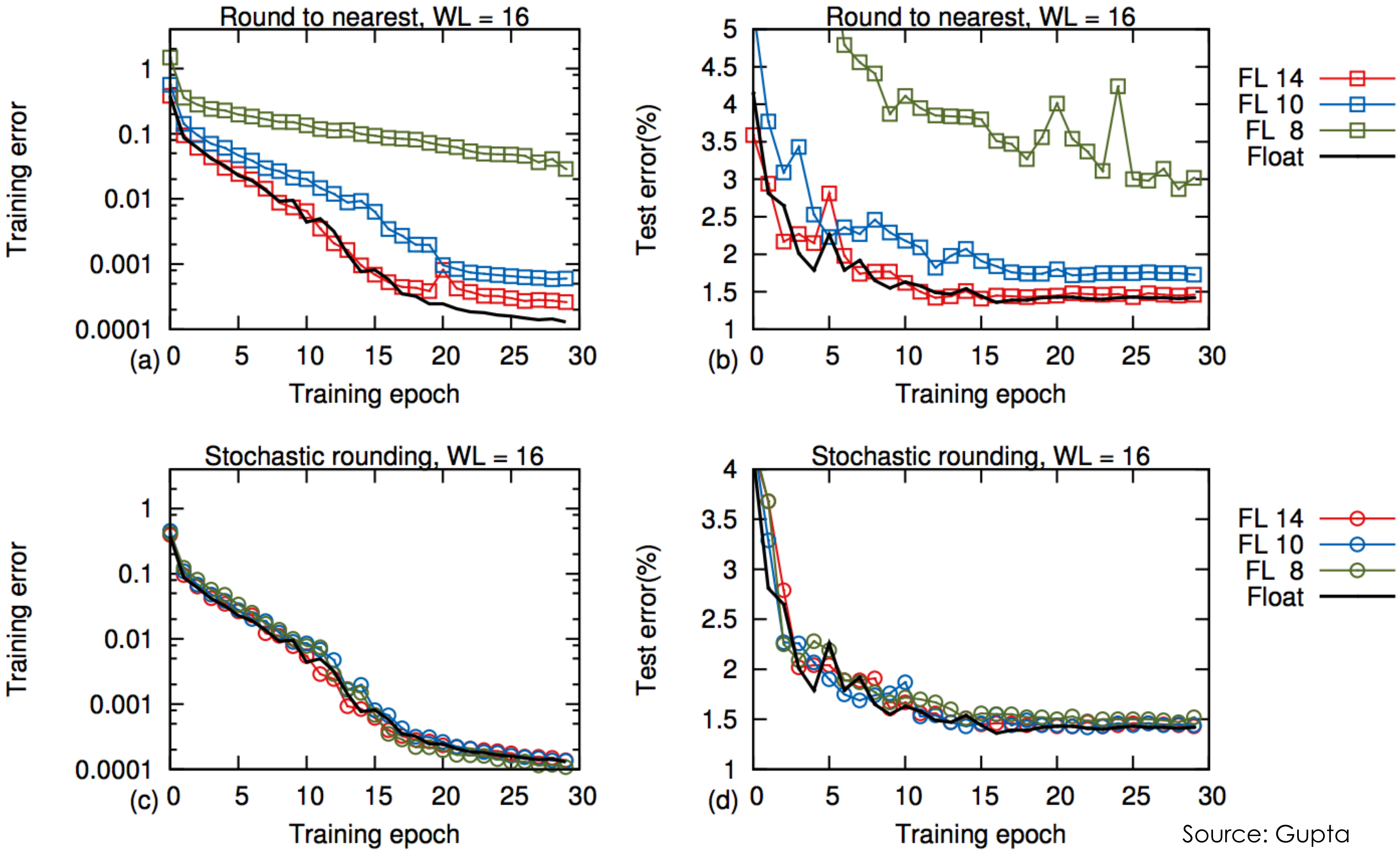


32-bit Single-Precision Floating-point Number



64-bit Double-Precision Floating-point Number

Low Precision Training

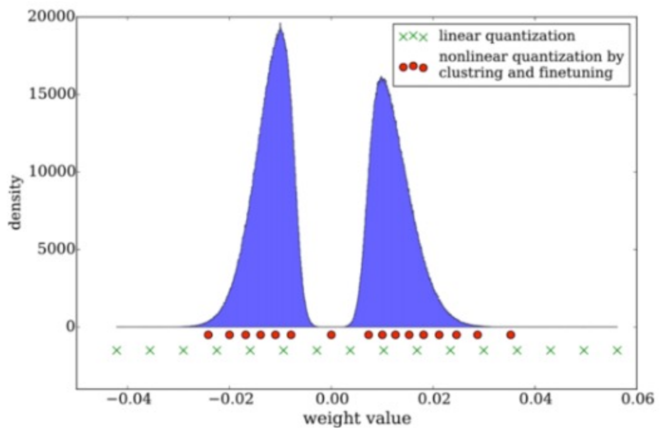


Source: Gupta

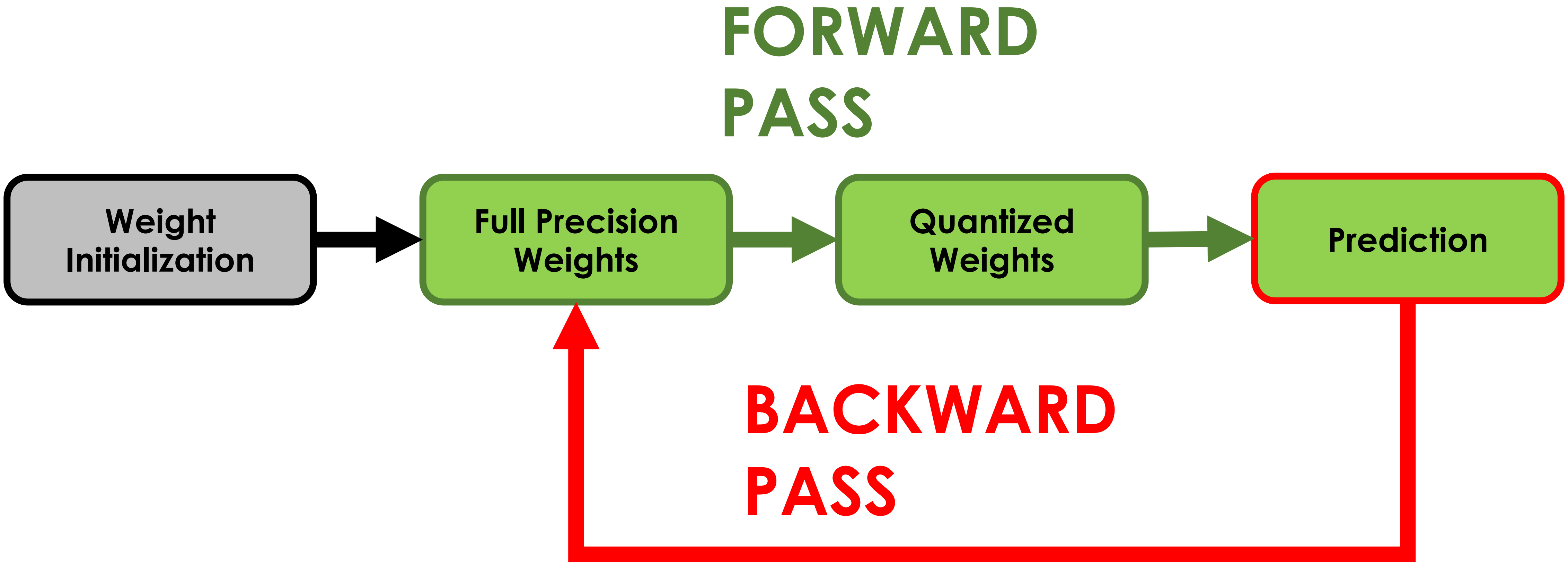
Quantization Implementation

- Linear vs. Non-linear
  - Linear – representable space is divided equally "X"
  - Non-linear – representable space is divided un-equally "O"

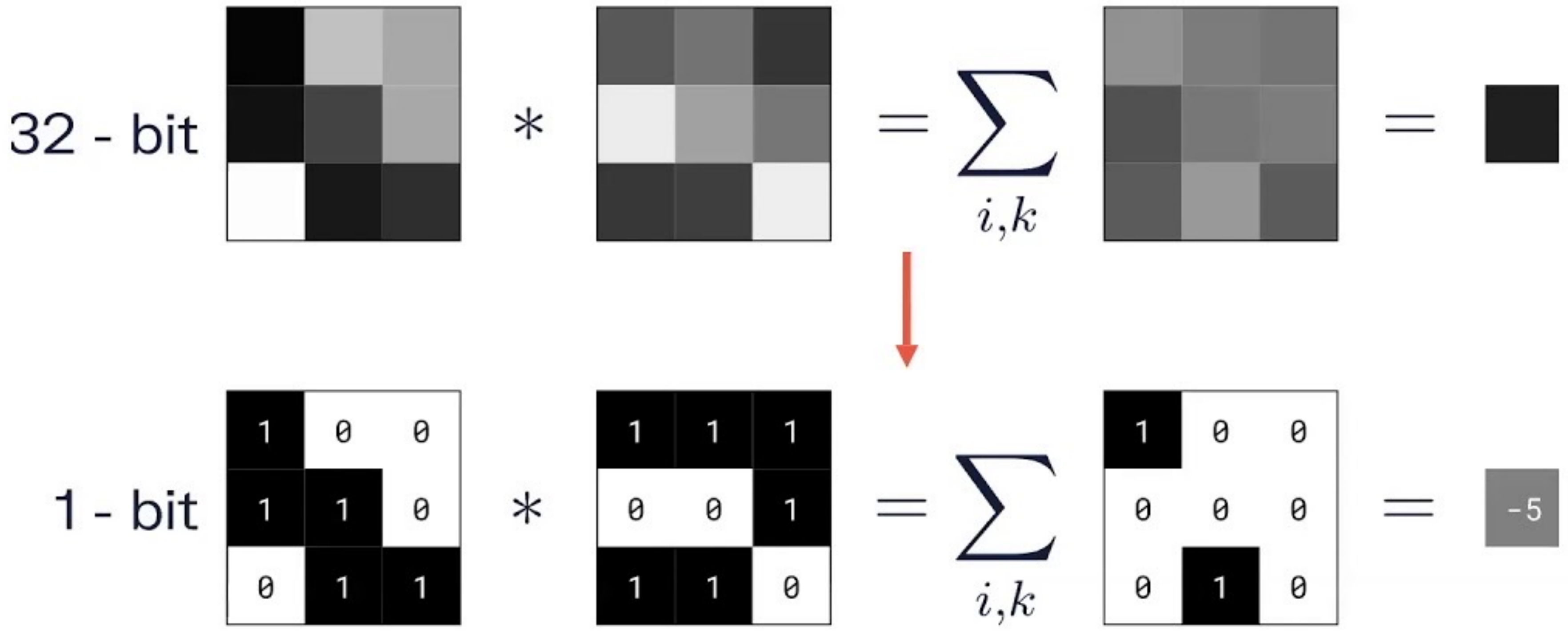
- Round to nearest vs. Stochastic rounding
  - Round to nearest – each number is represented by the closest representable number
  - Stochastic rounding – each number is represented by the closest higher or lower value with equal probabilities



Training Quantized Models



Binary: Extreme Quantization



Source: Geiger

Space Reduction

32x

Compute Reduction

No FP Ops!

Roadmap for Today

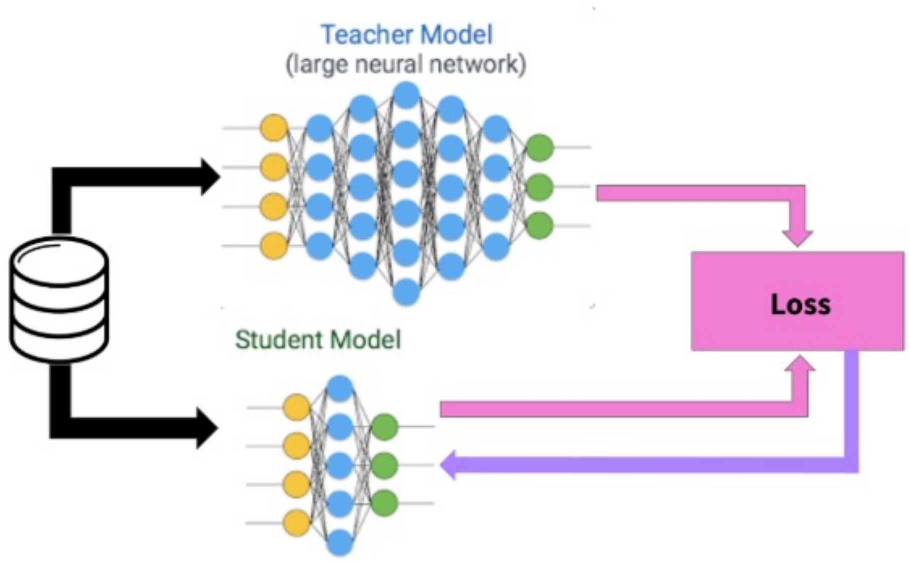
Architecture (*Re*-)Design

1. Early Evolution; Decomposition as a tool
2. Efficient Architectures; Example: MobileNet

Architecture Compression

3. Parameter and Channel Pruning
4. Parameter Quantization
5. **Knowledge Distillation**
6. Compression pipelines

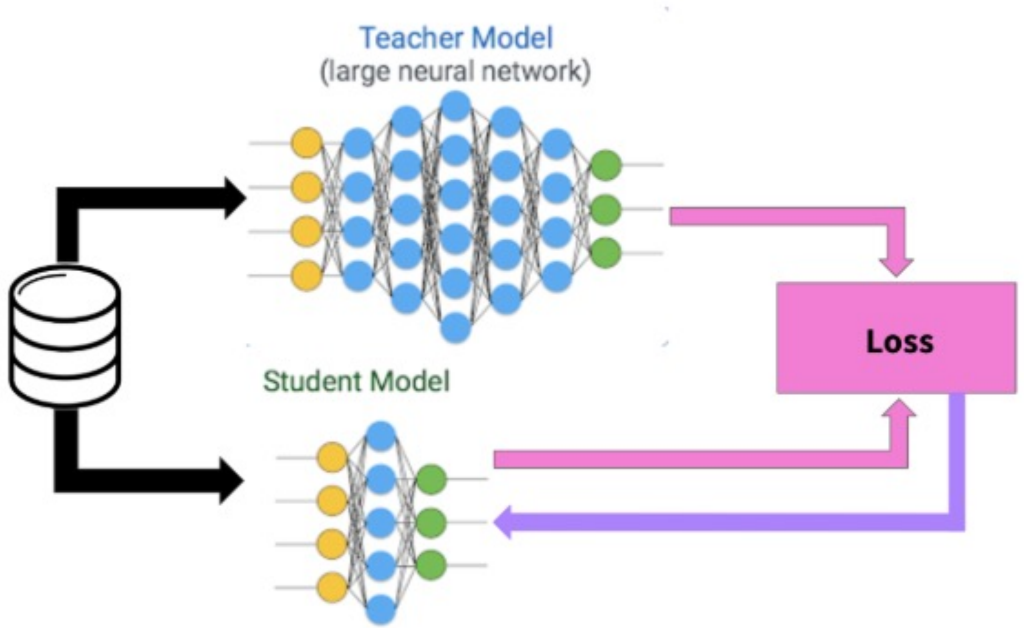
Knowledge Distillation



Distillation of a large model into a small one is performed in a teacher-student setup whereby the student network is trained to **replicate the teacher's raw output** (logits).

Knowledge Distillation

- **Knowledge distillation does not require actual data for the student**'s training since the goal is for it to match the teacher's output not the implied labels.
- Thus, students can be trained on random inputs just as well as on the original data! **The teacher provides the targets in real time.**



Knowledge Distillation

#	Model	SST-2 Acc	QQP F_1/Acc	MNLI-m Acc	MNLI-mm Acc
1	$\text{BERT}_{\text{LARGE}}$ (Devlin et al., 2018)	94.9	72.1/89.3	86.7	85.9
2	$\text{BERT}_{\text{BASE}}$ (Devlin et al., 2018)	93.5	71.2/89.2	84.6	83.4
3	OpenAI GPT (Radford et al., 2018)	91.3	70.3/88.5	82.1	81.4
4	BERT ELMo baseline (Devlin et al., 2018)	90.4	64.8/84.7	76.4	76.1
5	GLUE ELMo baseline (Wang et al., 2018)	90.4	63.1/84.3	74.1	74.5
6	Distilled $\text{BiLSTM}_{\text{SOFT}}$	**90.7**	**68.2/88.1**	**73.0**	**72.6**
7	BiLSTM (our implementation)	86.7	63.7/86.2	68.7	68.3
8	BiLSTM (reported by GLUE)	85.9	61.4/81.7	70.3	70.8
9	BiLSTM (reported by other papers)	$87.6^{\dagger}$	– /$82.6^{\ddagger}$	66.9^{*}	66.9^{*}

Source: Tang

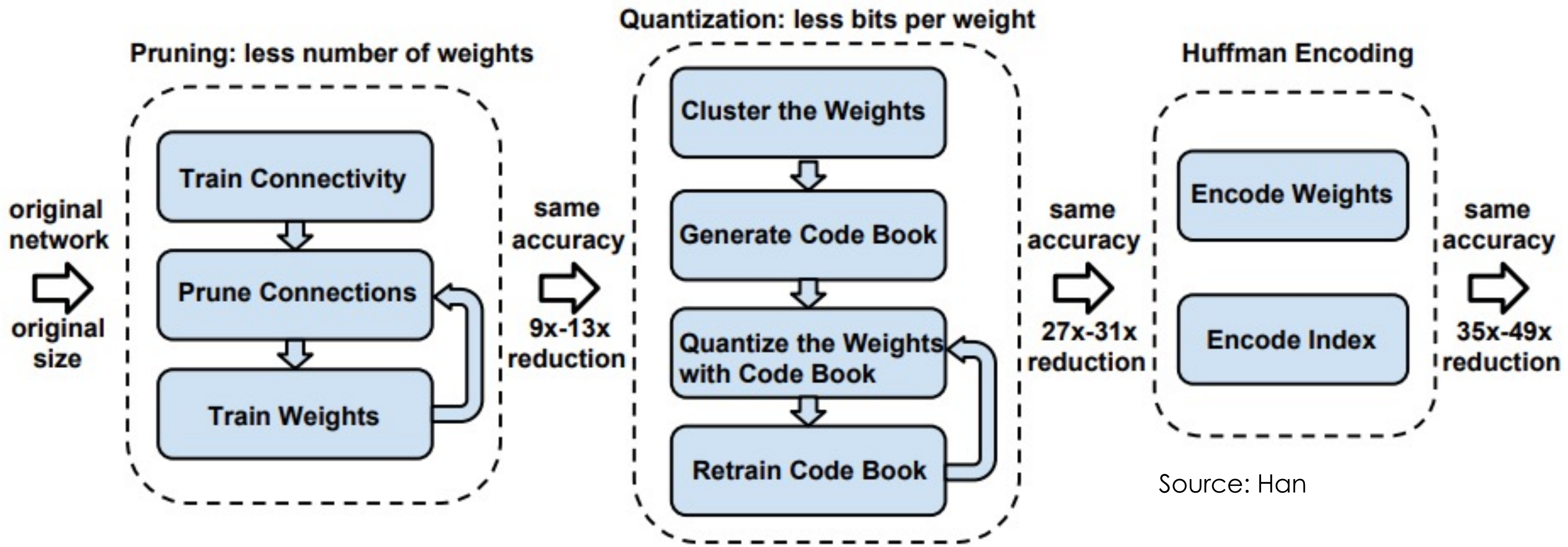
Achieve comparable results with ELMo, while using roughly 100 times fewer parameters and 15 times less inference time.

Roadmap for Today

Architecture (*Re*-)Design
1. Early Evolution; Decomposition as a tool
2. Efficient Architectures; Example: MobileNet

Architecture Compression
3. Parameter and Channel Pruning
4. Parameter Quantization
5. Knowledge Distillation
6. **Compression pipelines**

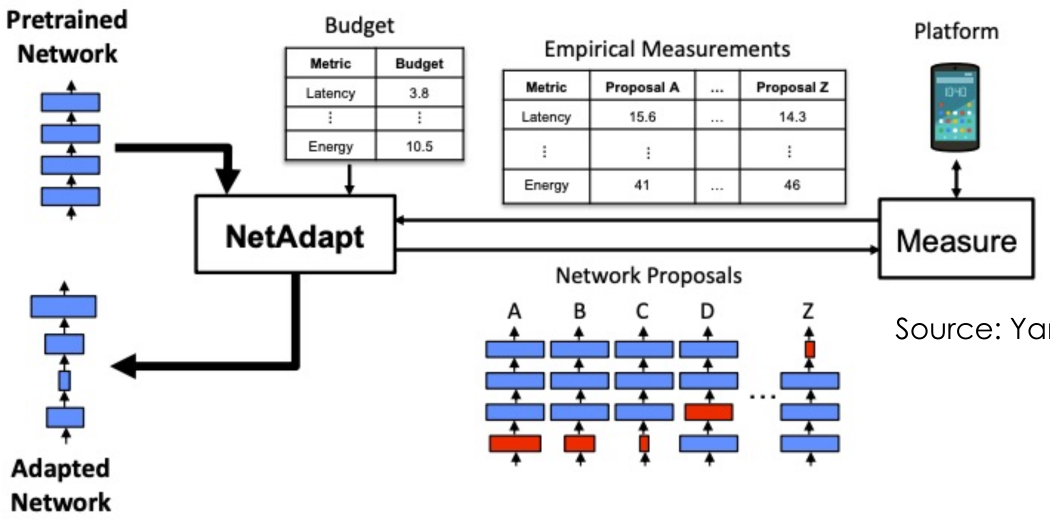
Basic Compression Pipeline



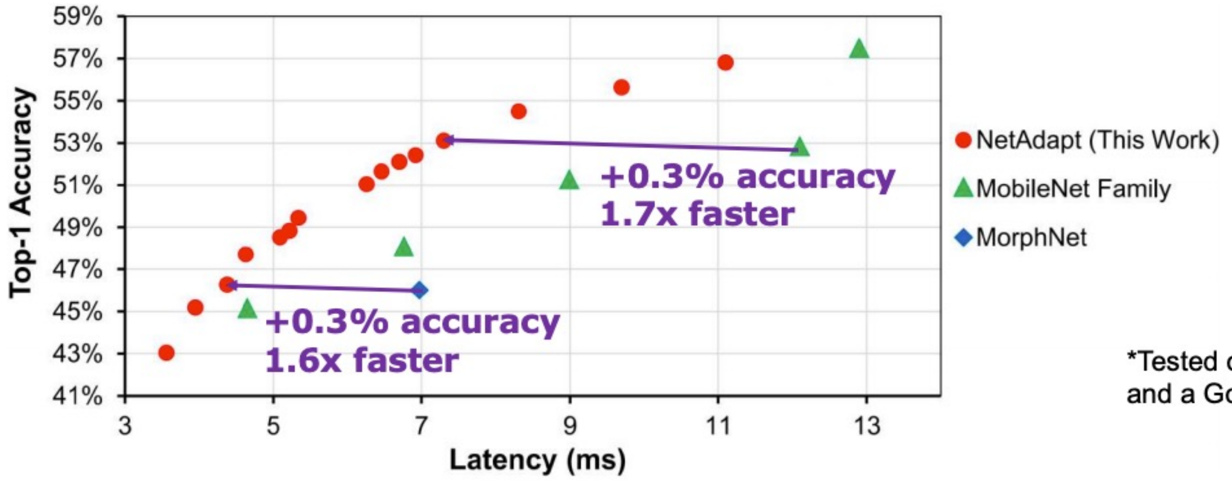
Source: Han

Resource-aware Pipeline

- **Automatically adapt DNN** to a mobile platform to reach a target latency or energy budget
- Use **empirical measurements** to guide optimization (avoid modeling of tool chain or platform architecture)
- Requires **very few hyperparameters** to tune



Source: Yang



*Tested on the ImageNet dataset and a Google Pixel 1 CPU

Summary of the Day

Architecture (*Re*-)Design

1. Early Evolution; Decomposition as a tool
2. Efficient Architectures; Example: MobileNet

Architecture Compression

3. Parameter and Channel Pruning
4. Parameter Quantization
5. Knowledge Distillation
6. Compression pipelines