Accelerator Architectures for Machine Learning (AAML)

Lecture 3: Quantization

Tsung Tai Yeh
Department of Computer Science
National Yang-Ming Chiao Tung University

Acknowledgements and Disclaimer

- Slides was developed in the reference with
  Joel Emer, Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, ISCA 2019 tutorial
  Efficient Processing of Deep Neural Network, Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, Joel Emer, Morgan and Claypool Publisher, 2020
  Yakun Sophia Shao, EE290-2: Hardware for Machine Learning, UC Berkeley, 2020
  CS231n Convolutional Neural Networks for Visual Recognition, Stanford University, 2020
- 6.5940, TinyML and Efficient Deep Learning Computing, MIT
- NVIDIA, Precision and performance: Floating point and IEEE 754 Compliance for NVIDIA GPUs, TB-06711-001_v8.0, 2017

Outline

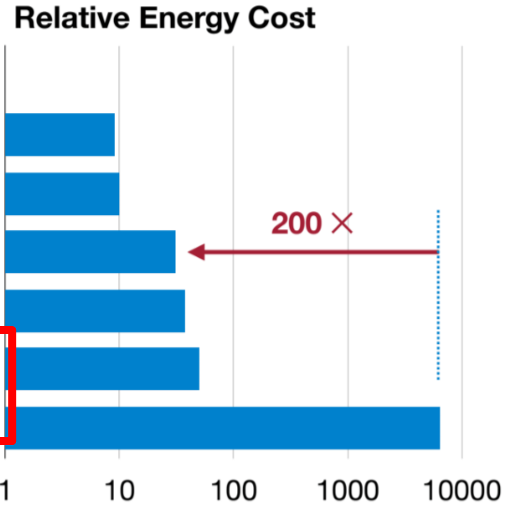
- K-Means-based Quantization
- Linear Quantization
- Binary and Ternary Quantization

Memory is Expensive !!

- **Data movement -> Move memory reference -> More energy**

Operation	Energy [pJ]
32 bit int ADD	0.1
32 bit float ADD	0.9
32 bit Register File	1
32 bit int MULT	3.1
32 bit float MULT	3.7
32 bit SRAM Cache	5
32 bit DRAM Memory	640

Rough Energy Cost For Various Operations in 45nm 0.9V



1  = 200 ×+

This image is in the public domain

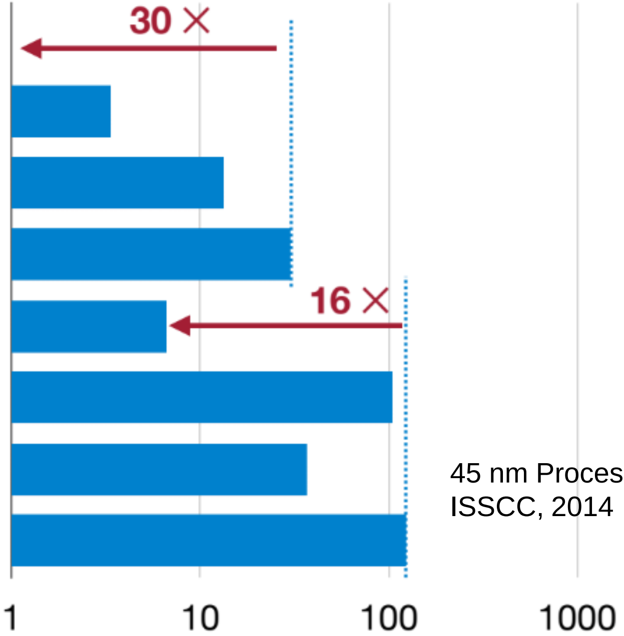
Computing's Energy Problem (and What We Can Do About it) [Horowitz, M., IEEE ISSCC 2014]

Low Bit-Width Operations are Cheap

- **Less Bit-Width -> Less energy**

Operation	Energy [pJ]
8 bit int ADD	0.03
32 bit int ADD	0.1
16 bit float ADD	0.4
32 bit float ADD	0.9
8 bit int MULT	0.2
32 bit int MULT	3.1
16 bit float MULT	1.1
32 bit float MULT	3.7

Rough Energy Cost For Various Operations in 45nm 0.9V



45 nm Process, Horowitz, ISSCC, 2014

Energy and Area Cost

Could we make the deep learning efficient by lowering the precision of data ?

Operation	**Energy (pJ)**	**Area(um^2)**
8b Add	0.03	36
16b Add	0.05	67
32b Add	0.1	137
16b FP Add	0.4	1360
32b FP Add	0.9	4184
16b FP Mult	1.1	1640
32b FP Mult	3.7	7700
32b SRAM Read (8KB)	5	
32b DRAM Read	640	

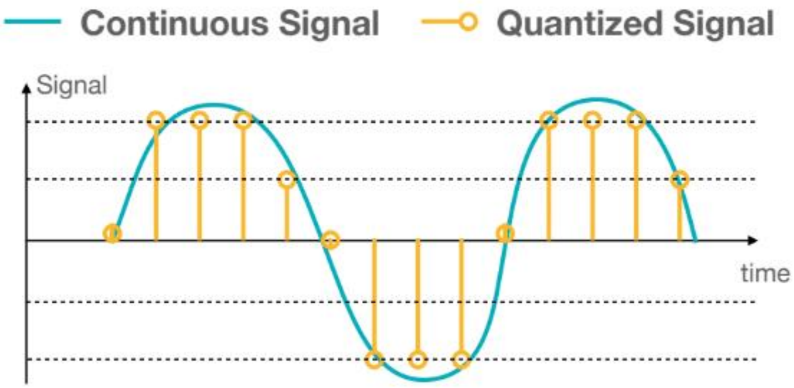
4.7X

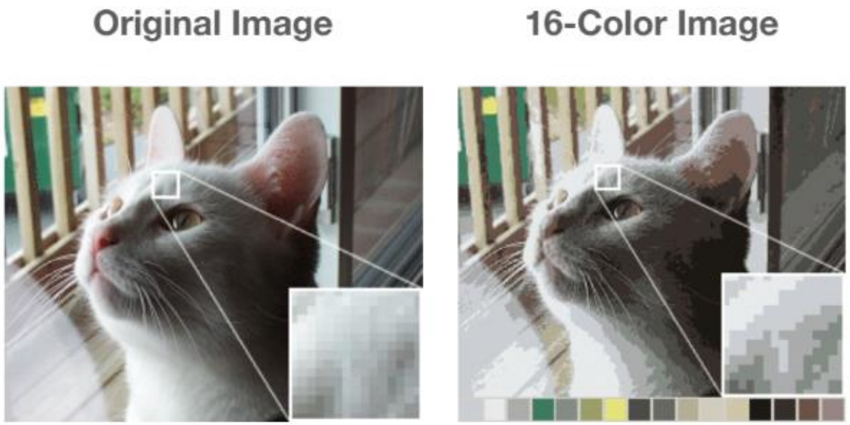
173X

45 nm Process, Horowitz, ISSCC, 2014

What is Quantization ?

- **Quantization**
  - The process of **constraining an input** from **a continuous or large set of values** to **a discrete set**

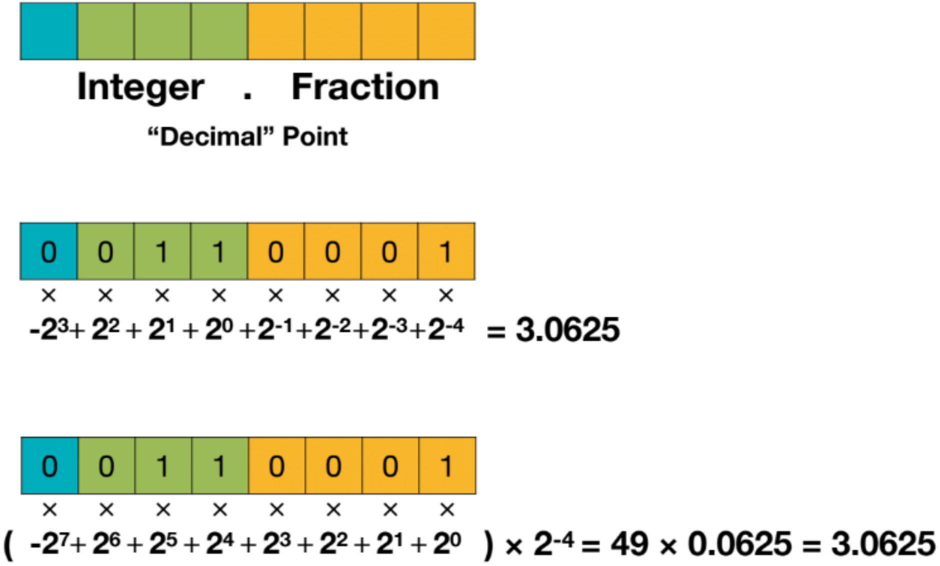




Images are in the public domain.

Numeric Data Types

- Fixed-point number



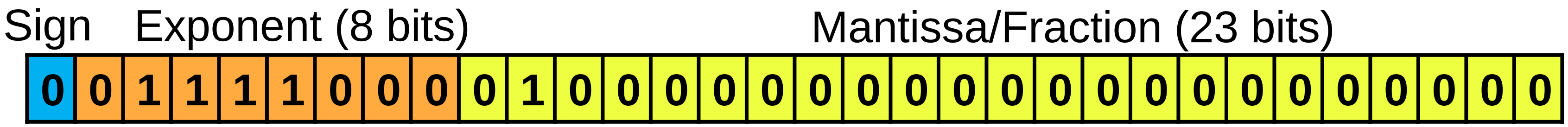
$$-2^3+2^2+2^1+2^0+2^{-1}+2^{-2}+2^{-3}+2^{-4} = 3.0625$$

$$(-2^7+2^6+2^5+2^4+2^3+2^2+2^1+2^0) \times 2^{-4} = 49 \times 0.0625 = 3.0625$$

(using 2's complement representation)

IEEE 765 Single Precision Float Point

- **Sign** determines the sign of the number
- **Exponent** (8 bit) represent -127 (all 0s) and +128 (all 1s)
- **Significand** (23 fraction bits), total precision is 24 bits (23 + 1 implicit leading bit) $\log_{10}(2^{24}) \approx 7.225$ digital bit

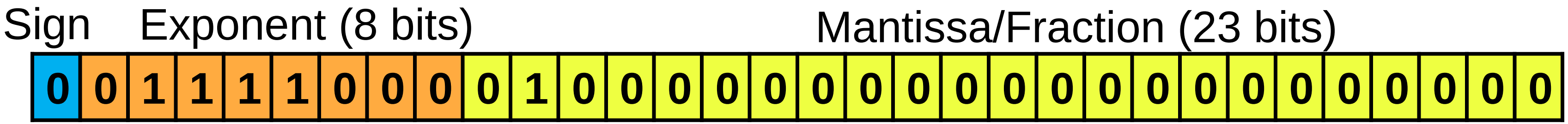


Sign Exponent (8 bits) Mantissa/Fraction (23 bits)

0	0	1	1	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

$$value = (-1)^{sign} \times 2^{(e-127)} \times (1 + \sum_{i=1}^{23} b_{(23-i)} 2^{-i})$$

IEEE 765 FP32 Case Study 1



0	0	1	1	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

$$value = (-1)^{sign} \times 2^{(e-127)} \times (1 + \sum_{i=1}^{23} b_{(23-i)} 2^{-i})$$

Sign = b31 = 0 ; $(-1)^0 = 1$

e =120; $2^{(120 - 127)} = 2^{-7}$

$$1.b_{22}b_{21}...b_0 = (1 + \sum_{i=1}^{23} b_{(23-i)} 2^{-i}) = 1 + 2^{-2} = 1.25$$

Value = 1 x 2^{-7} x 1.25 = 0.009765625

Numeric Data Type

- **Question**: What is the decimal **“11.375”** in FP32 format ?

11.375
= 11 + 0.375
= $(1011)_2 + (0.011)_2$
= $(1011.011)_2$
= $(1.011011)_2 \times 2^3$

$0.375 \times 2 = 0.750 = 0 + 0.750 \Rightarrow b_{-1} = 0$
$0.750 \times 2 = 1.500 = 1 + 0.500 \Rightarrow b_{-2} = 1$
$0.500 \times 2 = 1.000 = 1 + 0.000 \Rightarrow b_{-3} = 1$

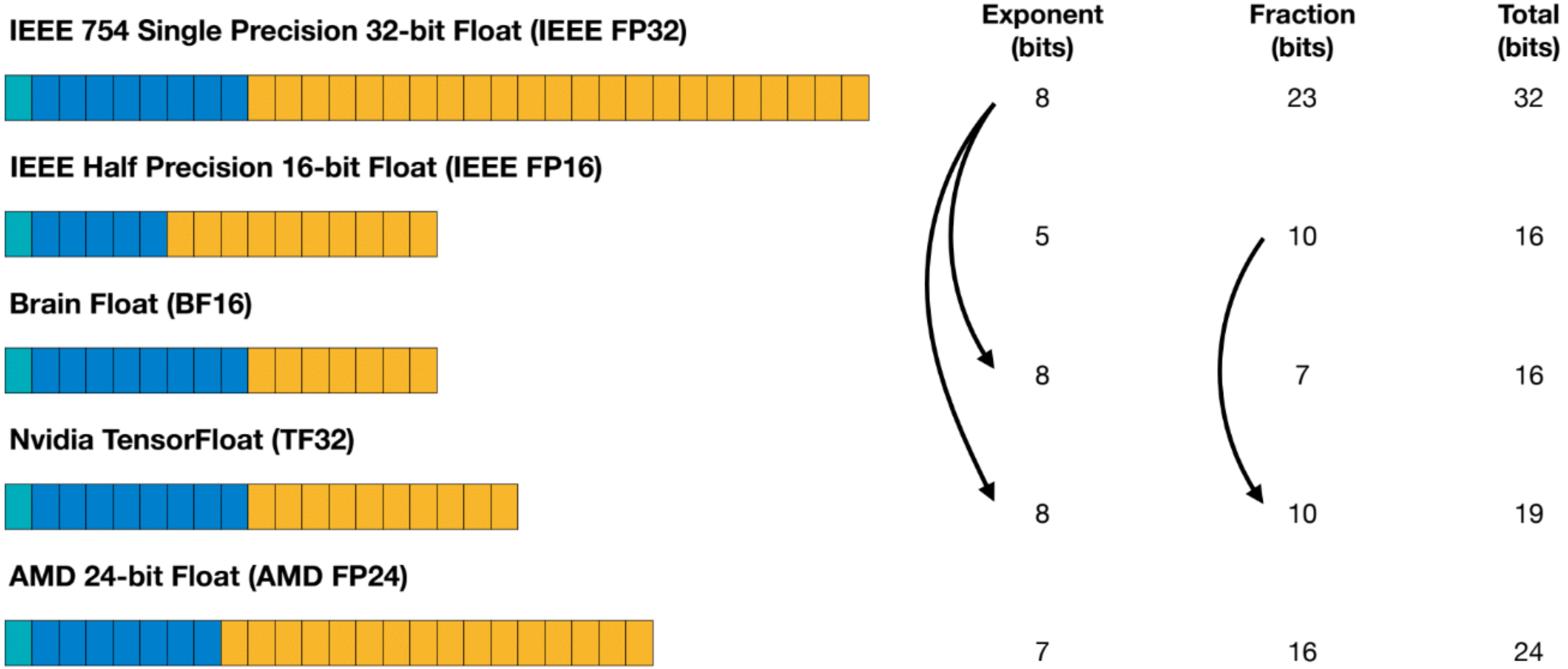
- The exponent is 3 and biased form

= (3 + 127) = 130 = 1000 0010

Sign	Exponent (8 bits)	Mantissa/Fraction (23 bits)
0	1 0 0 0 0 0 1 0	0 1 1 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

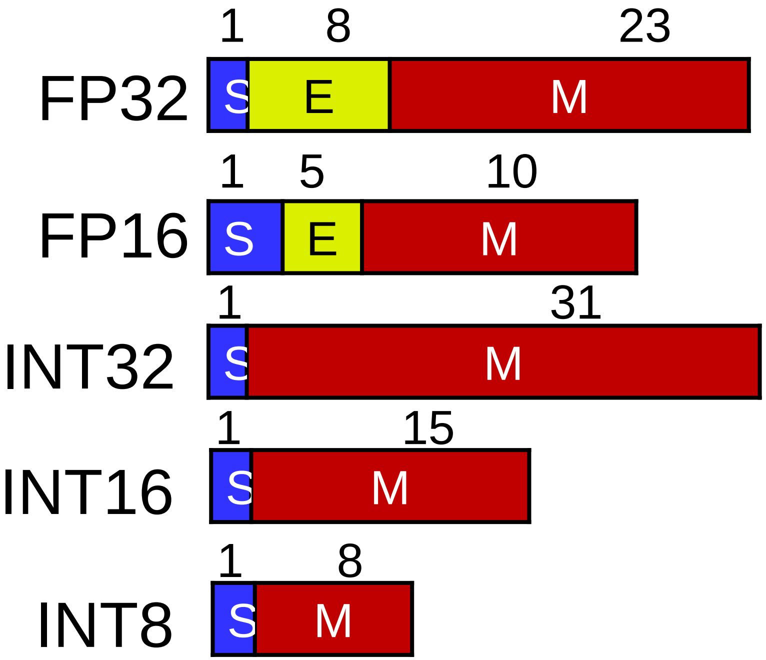
Floating-Point Number

- **Exponent Width -> Range; Fraction Width-> Precision**



	Exponent (bits)	Fraction (bits)	Total (bits)
IEEE 754 Single Precision 32-bit Float (IEEE FP32)	8	23	32
IEEE Half Precision 16-bit Float (IEEE FP16)	5	10	16
Brain Float (BF16)	8	7	16
Nvidia TensorFloat (TF32)	8	10	19
AMD 24-bit Float (AMD FP24)	7	16	24

Number Representation



Range

1.2E-38 to 3.4E+38

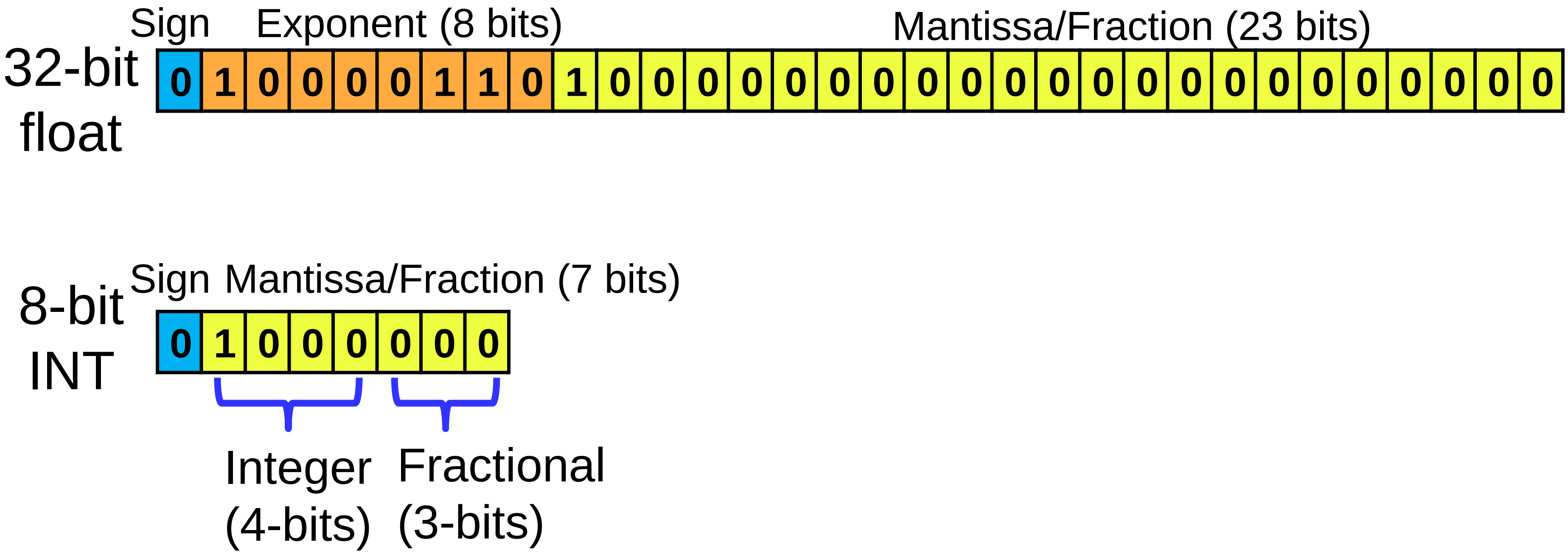
6.1E-5 to 6.6E+4

2147483648 to 2147483647

−32,768 to 32,767

-128 ~ 127

Reduced Bit Width



32-bit float

Sign | Exponent (8 bits) | Mantissa/Fraction (23 bits)

0	1	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

8-bit INT

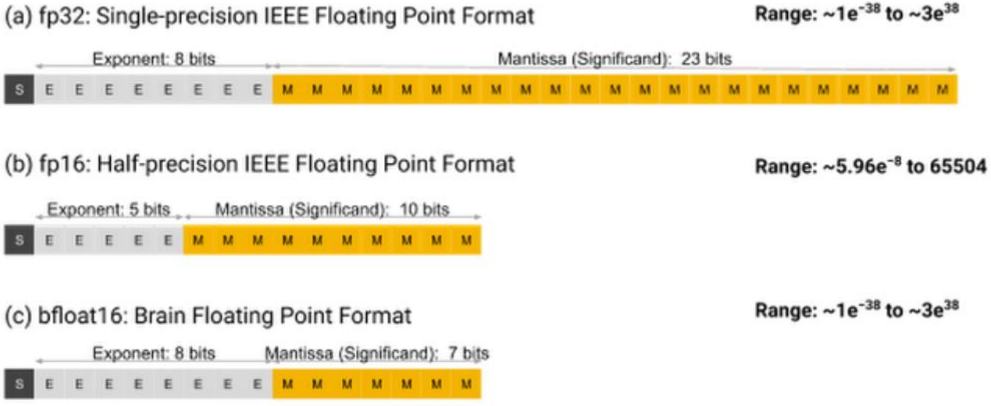
Sign | Mantissa/Fraction (7 bits)

0	1	0	0	0	0	0	0

Integer (4-bits) | Fractional (3-bits)

FP32 vs FP16 vs BF16

- **FP32 – single precision**
  - With 6-9 significant decimal digits precision
- **FP16 – half precision**
  - Uses in some neural network applications
  - With 4 significant decimal digits precision
- **BF16**
  - A truncated FP32
  - Allow for fast conversion to and from an FP32
  - With 3 significant decimal digits


(a) fp32: Single-precision IEEE Floating Point Format **Range: ~1e^{-38} to ~3e^{38}**

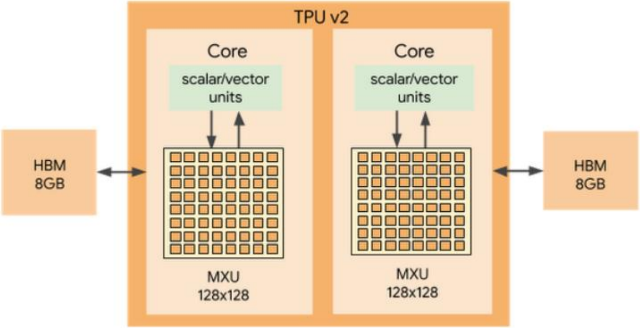
(b) fp16: Half-precision IEEE Floating Point Format **Range: ~5.96e^{-8} to 65504**

(c) bfloat16: Brain Floating Point Format **Range: ~1e^{-38} to ~3e^{38}**

Format	Bits	Exponent	Fraction
FP32	32	8	23
FP16	16	5	10
BF16	16	8	7

https://cloud.google.com/blog/products/ai-machine-learning/bfloat16-the-secret-to-high-performance-on-cloud-tpus

Choosing bFloat16

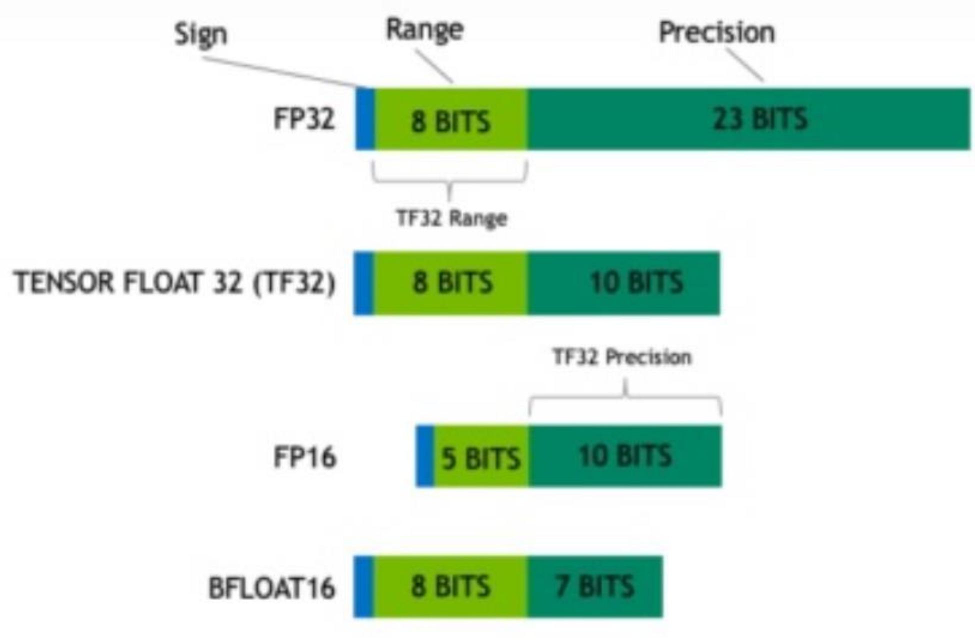


- **Motivation**
  - The physical size of a hardware multiplier scales with the square of the mantissa width
  - Mantissa bit length – FP32: 23, FP16: 10, BF16: 7
- **BF16**
  - 8 X smaller than an FP32 multiplier
  - Has the same exponent size as FP32
  - No require special handling (loss scaling) in the FP16 conversion
  - XLA compiler’s automatic format conversion
  - In side the MXU, multiplications are performed in BF16 format
  - Accumulations are performed in full FP32 precision

https://cloud.google.com/blog/products/ai-machine-learning/bfloat16-the-secret-to-high-performance-on-cloud-tpus

Nvidia’s TF32

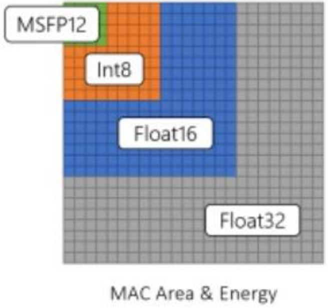
- **Nvidia’s TF32**
  - 19-bit (BF19)
  - 1-bit sign, 8-bit exponent 10-bit fraction
  - Fuse BF16 and FP16
    - BF16: 8-bit exponent +
    - FP16: 10-bit fraction
  - Nvidia A100 Tensor Core
    - TF32: 156 TFLOPS
    - FP16/BF16: 312 TFLOPS



https://zhuanlan.zhihu.com/p/449857213

Microsoft MSFP

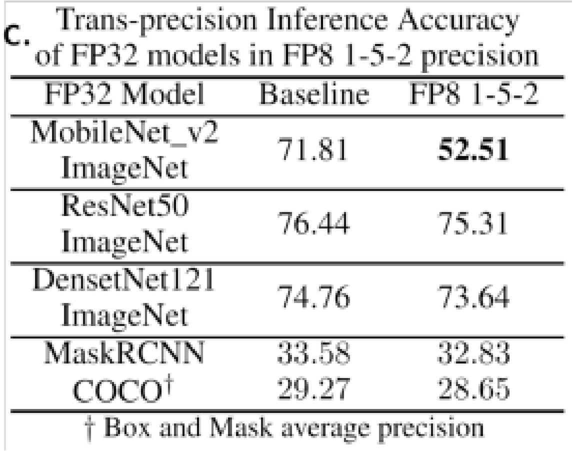
- **Microsoft MSFP**
  - Used in Brainwave FPGA
  - 8-bit shared exponent
  - 1-bit sign, 3-bit fraction
  - A group of INT4 vector shares 8-bit exponent



https://zhuanlan.zhihu.com/p/449857213

FP8 and Tesla CFloat

- **FP8 (1-5-2)**
  - Large loss in MobileNet v2
  - Hybrid FP8 (HFP8)
    - Use FP(1-4-3) in forward
    - Use FP(1-5-2) in backward
- **Tesla Dojo Cfloat (configurable float)**
  - Configurable exponent and mantissa
  - Use software to choose appropriate Cfloat format
    - CF16
    - CF8 (1-4-3), CF8 (1-5-2)



c. Trans-precision Inference Accuracy of FP32 models in FP8 1-5-2 precision

FP32 Model	Baseline	FP8 1-5-2
MobileNet_v2 ImageNet	71.81	**52.51**
ResNet50 ImageNet	76.44	75.31
DensetNet121 ImageNet	74.76	73.64
MaskRCNN COCO†	33.58 29.27	32.83 28.65

† Box and Mask average precision

https://proceedings.neurips.cc/paper/2019/file/65fc9fb4897a89789352e211ca2d398f-Paper.pdf

How to Determine Bit Width on DNN ?

- For accuracy, DNN operations decide bit width to achieve sufficient precision
- Which DNN operations affect the accuracy ?
    - **For inference:** weights, activations, and partial sums
    - **For training:** weights, activations, partial sums, gradients, and weight update

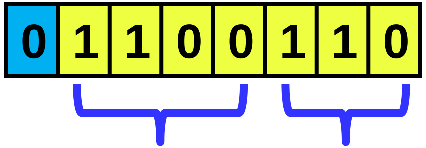
Dynamic Fixed Point

- Allow “**f**” to vary based on data type and layer
- In large layers, the outputs are the result of many accumulations
- The value of network parameters are much smaller than layer output -> **varying bit widths on parameters and outputs**

8-bit dynamic fixed

12.75

Sign Mantissa (7 bits)

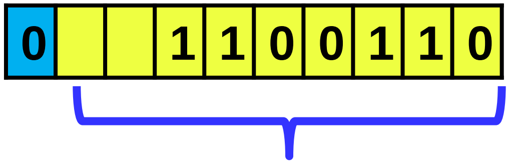


Integer ([7-f]-bits)

Fractional (f-bits)

S = 0, m = 102, f = 3

Sign Mantissa (7 bits)



0.19921875

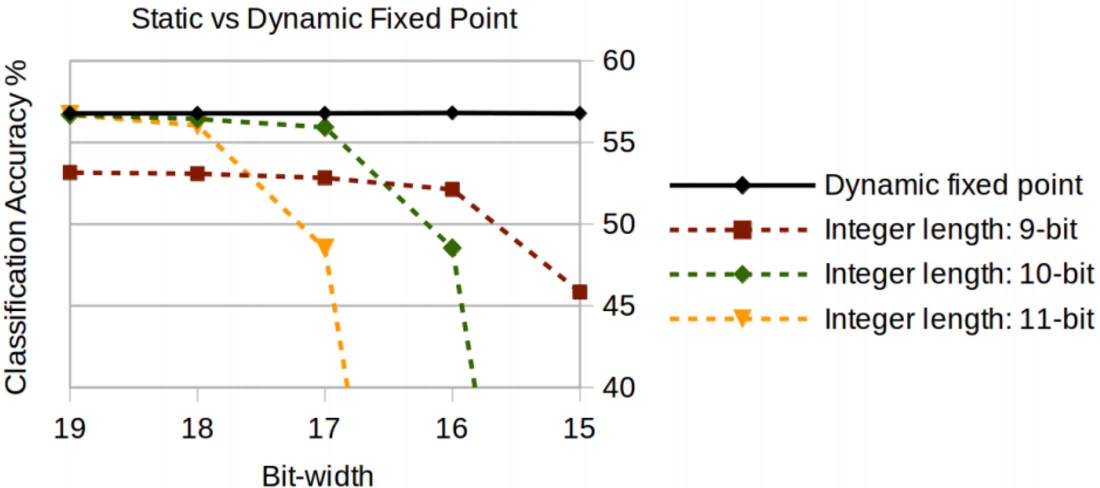
fractional (f-bits)

S = 0, m = 102, f = 9

Impact on Accuracy

- The accuracy drops in the small bit width when using static fixed point
- Stable accuracy variation is shown in dynamic fixed point (why ?)

Top-1 accuracy of CaffeNet on ImageNet



Gysel et al., Ristretto, ICLR 2016

Impact on Accuracy

- Small bit width cannot adapt to every DNN models very well (training)

	Layer outputs	CONV parameters	FC parameters	Fixed point accuracy
LeNet (Exp 1)	4-bit	4-bit	4-bit	99.0%
LeNet (Exp 2)	4-bit	2-bit	2-bit	98.8%
SqueezeNet	8-bit	8-bit	8-bit	57.1%
CaffeNet	8-bit	8-bit	8-bit	56.0%
GoogleNet	8-bit	8-bit	8-bit	66.6%

Gysel et al., Ristretto, ICLR 2016

Precision Varies from Layer to Layer

- Accuracy varies with the different bit widths in layers
- How to find out the best bit width in each layer while maintaining high accuracy ?

AlexNet

Error rate	**Bit per layer**
1%	10-8-8-8-8-8-6-4
2%	10-8-8-8-8-8-5-4
5%	10-8-8-8-7-7-5-3
10%	9-8-8-8-7-7-5-3

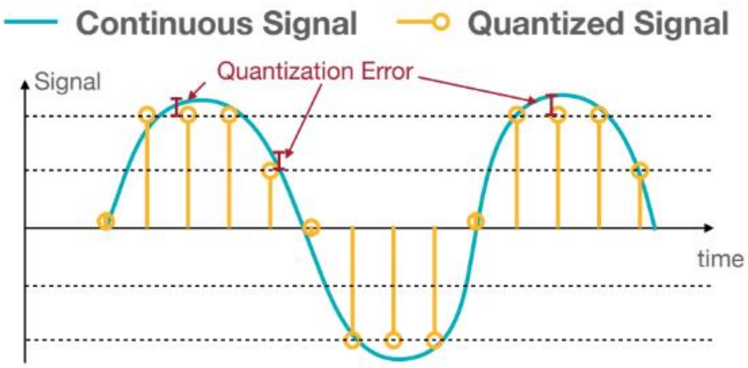
Judd et. al., ArXiv 2016

Takeaway Questions

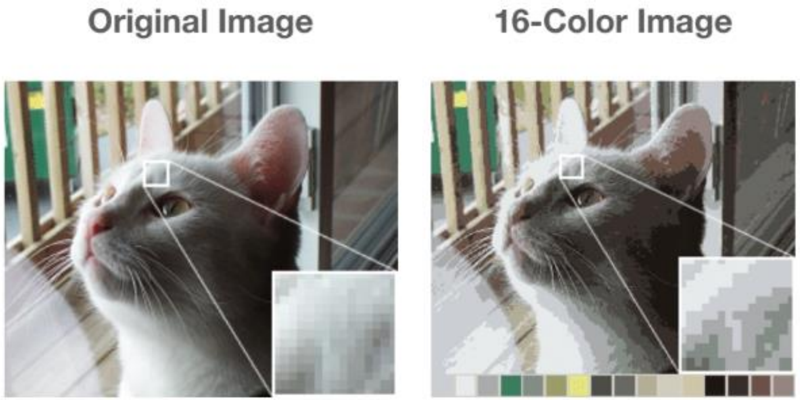
- What are advantages to use BF16 instead of FP16 ?
    - (A) Fast conversion from FP32
    - (B) Get more precise value
    - (C) Represent few different values
- What are benefits to use lower precision data type on neural network ?
    - (A) Reduce the latency of DNN models
    - (B) Save the memory space
    - (C) Lower the power consumption of the accelerator

What is Quantization ?

- **Quantization** is the process of **constraining an input from a continuous or large set of values** to **a discrete set**



The difference between an input value and its quantized value is referred to as quantization error.


Original Image

16-Color Image

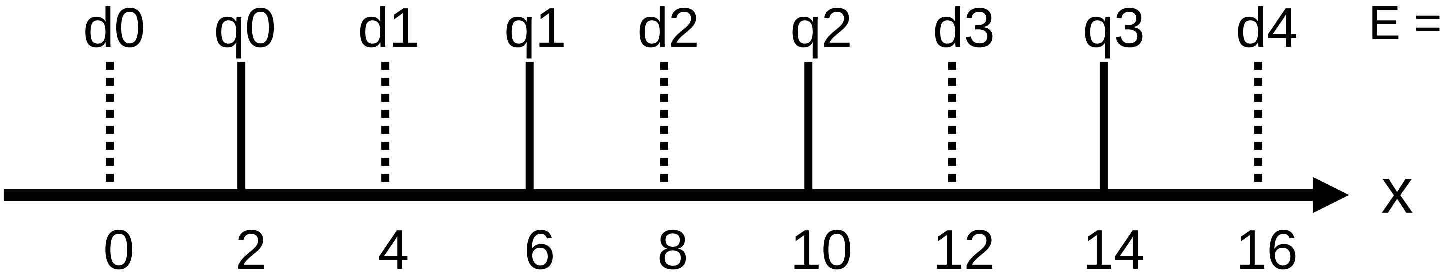
Images are in the public domain.

“Palettization”

Quantization [Wikipedia]

Data Quantization

- **Quantization**
  - Maps data from a full precision to reduced one
- **Quantization error**
  - Measures the average difference between the original full precision and quantized values

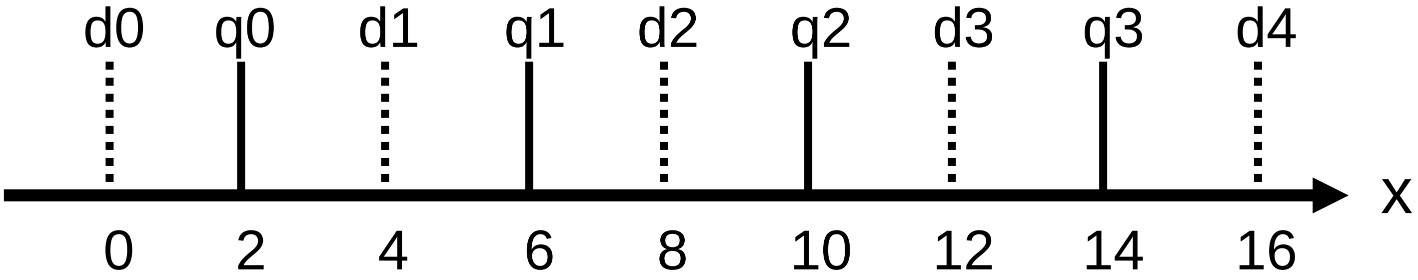


$E = [(x - x^*)^2]$

$x = 1, 3, 7, 8, 15$ → Uniform Quantization → $x^* = 2, 2, 6, 8, 14$

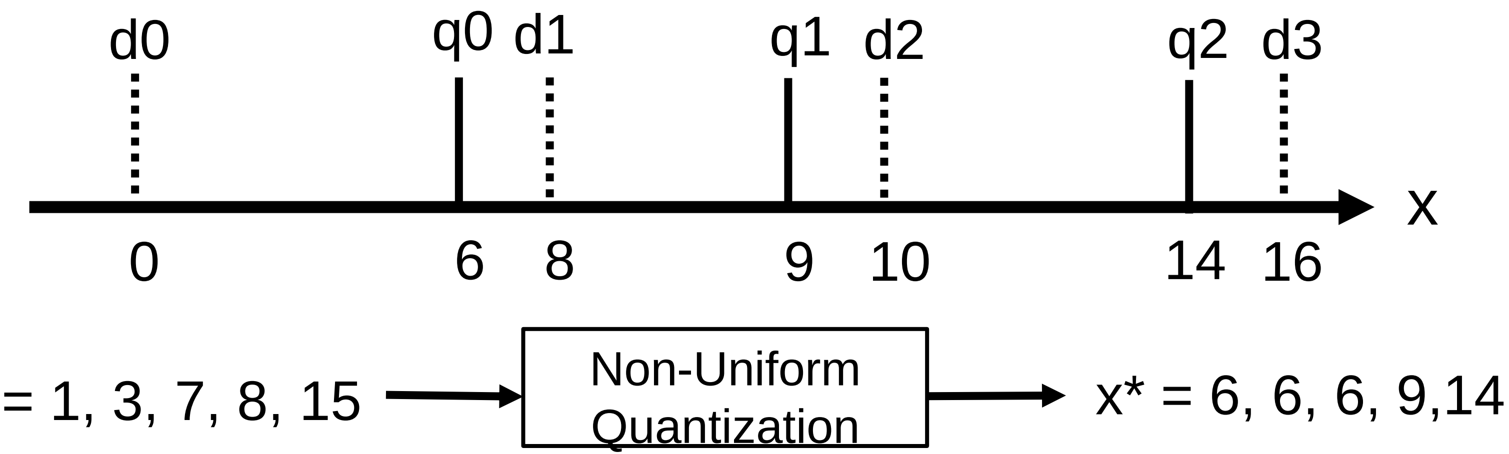
Types of Quantization

- **Uniform Quantization**
    - Quantized values are equally spaced out
    - x* can take on are {2, 6, 10, 14} with level = 4
    - Decision boundaries di are used to decide the quantization value that x should be mapped to



Types of Quantization

- **Non-uniform quantization**
    - Spacing can be computed e.g. logarithmic or with look-up-table
    - Fewer unique values can make weight sharing and compression

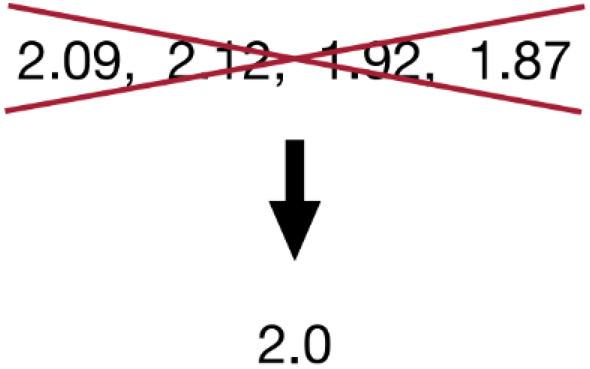


K-Means-based Weight Quantization

- **Storage**
  - Integer Weights; Floating-Point Codebook
- **Computation**
  - Floating-Point Arithmetic

weights
(32-bit float)

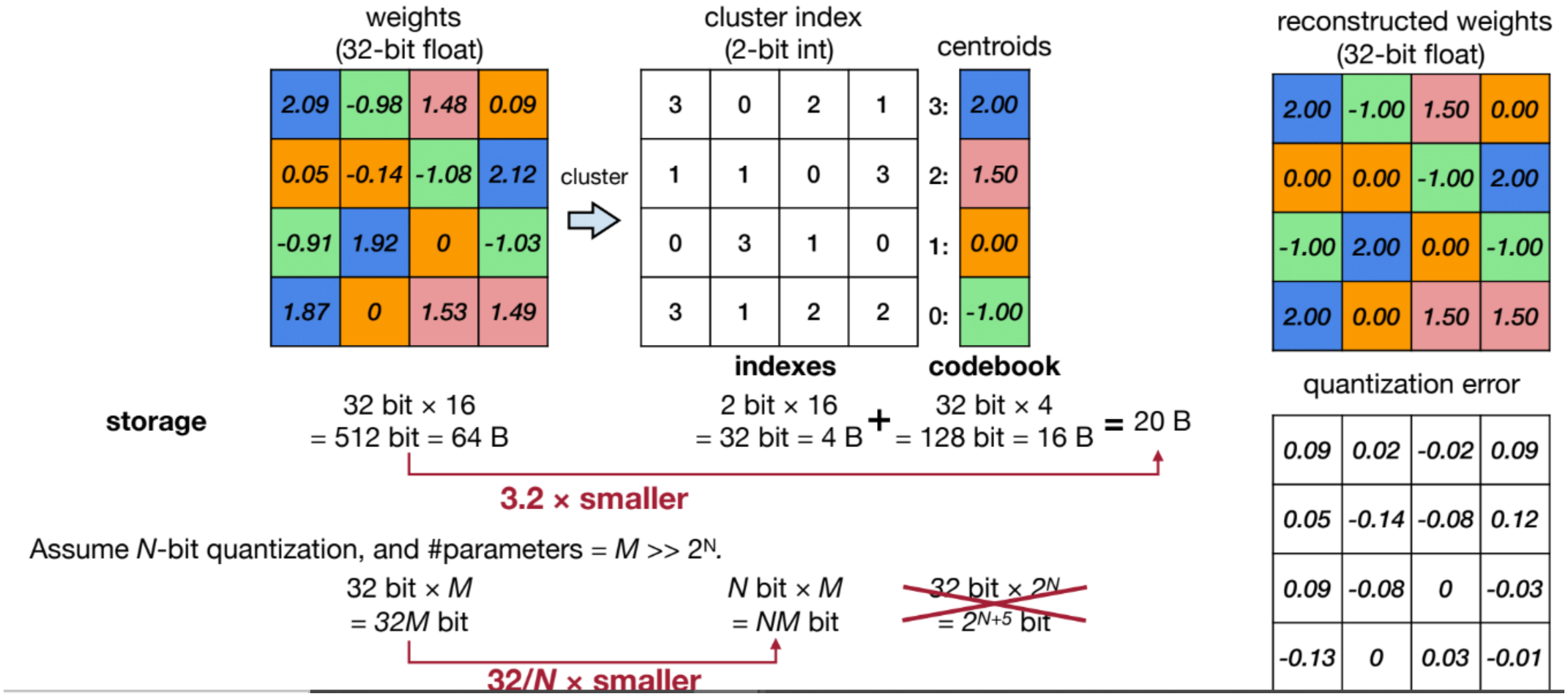
2.09	*-0.98*	*1.48*	*0.09*
0.05	*-0.14*	*-1.08*	*2.12*
-0.91	*1.92*	*0*	*-1.03*
1.87	*0*	*1.53*	*1.49*



~~2.09, 2.12, 1.92, 1.87~~

2.0

K-Means-based Weight Quantization



weights
(32-bit float)

2.09	-0.98	1.48	0.09
0.05	-0.14	-1.08	2.12
-0.91	1.92	0	-1.03
1.87	0	1.53	1.49

cluster index
(2-bit int)

3	0	2	1
1	1	0	3
0	3	1	0
3	1	2	2

indexes

centroids

3:	2.00
2:	1.50
1:	0.00
0:	-1.00

codebook

reconstructed weights
(32-bit float)

2.00	-1.00	1.50	0.00
0.00	0.00	-1.00	2.00
-1.00	2.00	0.00	-1.00
2.00	0.00	1.50	1.50

quantization error

0.09	0.02	-0.02	0.09
0.05	-0.14	-0.08	0.12
0.09	-0.08	0	-0.03
-0.13	0	0.03	-0.01

storage

32 bit × 16
= 512 bit = 64 B

2 bit × 16
= 32 bit = 4 B

\+

32 bit × 4
= 128 bit = 16 B

= 20 B

3.2 × smaller

Assume *N*-bit quantization, and #parameters = $M >> 2^N$.

32 bit × M
= $32M$ bit

N bit × M
= NM bit

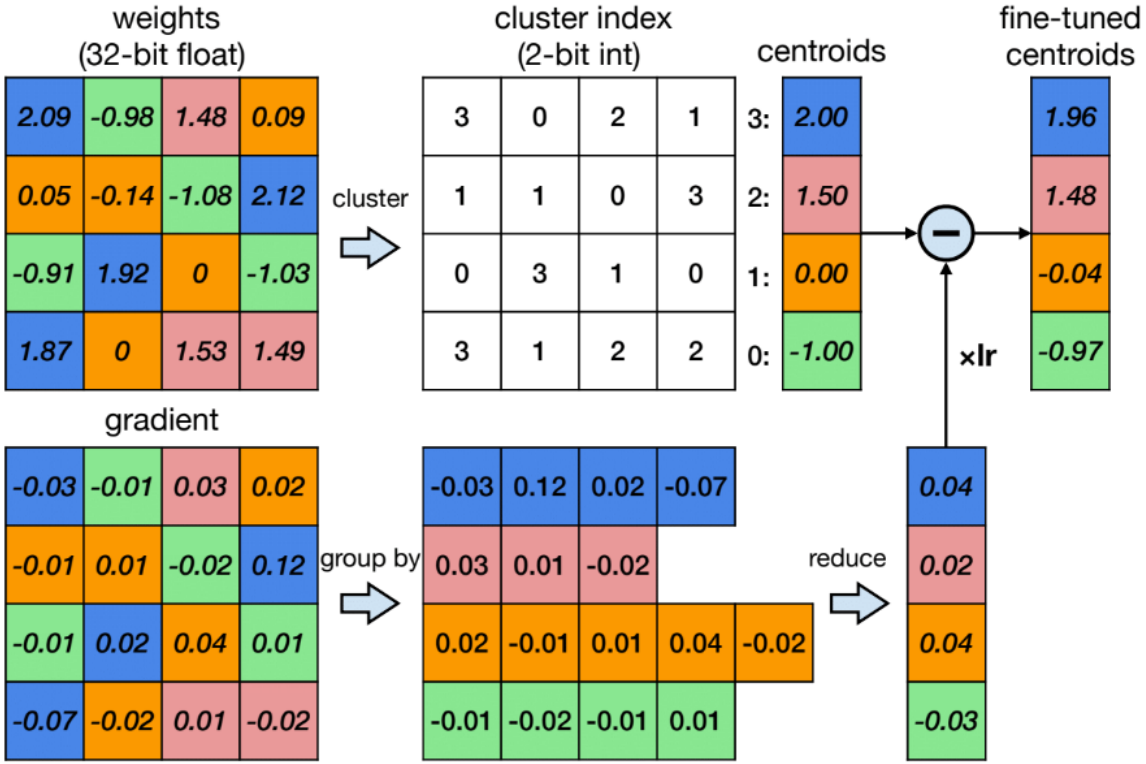
~~32 bit × 2^N~~
~~= 2^{N+5} bit~~

32/*N* × smaller

K-Means-based Weight Quantization

- **Fine-tuning**

  Quantized Weights



weights (32-bit float)

2.09	-0.98	1.48	0.09
0.05	-0.14	-1.08	2.12
-0.91	1.92	0	-1.03
1.87	0	1.53	1.49

cluster index (2-bit int)

3	0	2	1
1	1	0	3
0	3	1	0
3	1	2	2

	centroids	fine-tuned centroids
3:	2.00	1.96
2:	1.50	1.48
1:	0.00	-0.04
0:	-1.00	-0.97

gradient

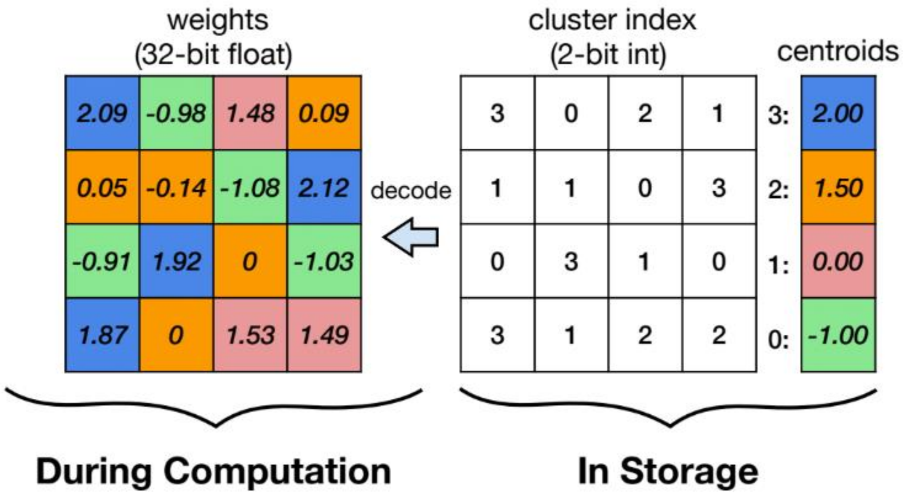
-0.03	-0.01	0.03	0.02
-0.01	0.01	-0.02	0.12
-0.01	0.02	0.04	0.01
-0.07	-0.02	0.01	-0.02

-0.03	0.12	0.02	-0.07	
0.03	0.01	-0.02		
0.02	-0.01	0.01	0.04	-0.02
-0.01	-0.02	-0.01	0.01	

0.04
0.02
0.04
-0.03

K-Means-based Weight Quantization

- Weights are decompressed using a lookup table during runtime inference
- Only saves storage cost of a neural network model
- All the computation and memory access are still floating-point



weights (32-bit float)

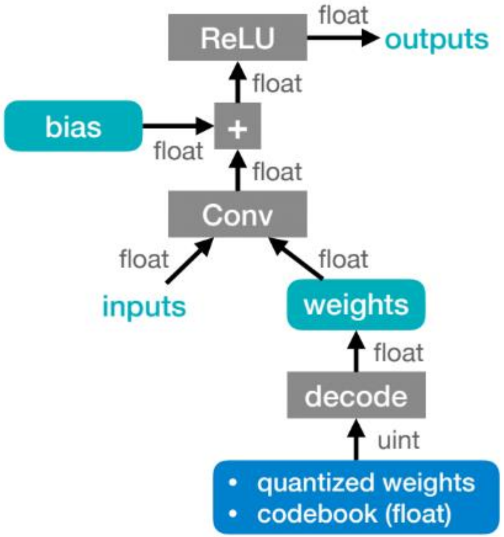
2.09	-0.98	1.48	0.09
0.05	-0.14	-1.08	2.12
-0.91	1.92	0	-1.03
1.87	0	1.53	1.49

cluster index (2-bit int)

3	0	2	1
1	1	0	3
0	3	1	0
3	1	2	2

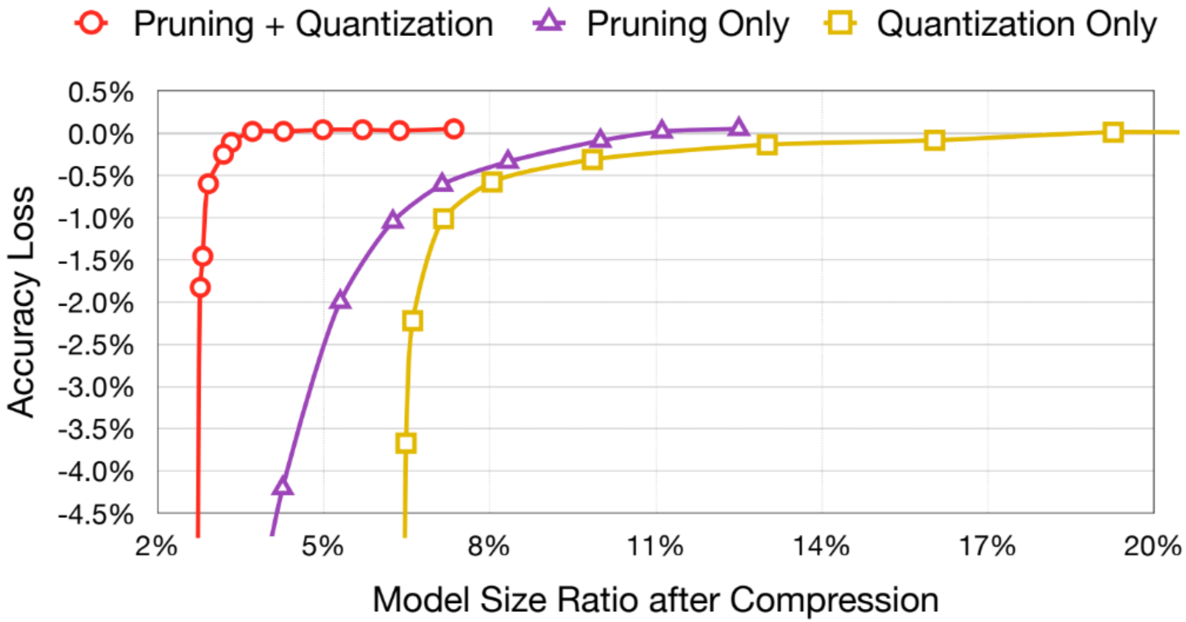
centroids

3:	2.00
2:	1.50
1:	0.00
0:	-1.00



K-Means-based Weight Quantization

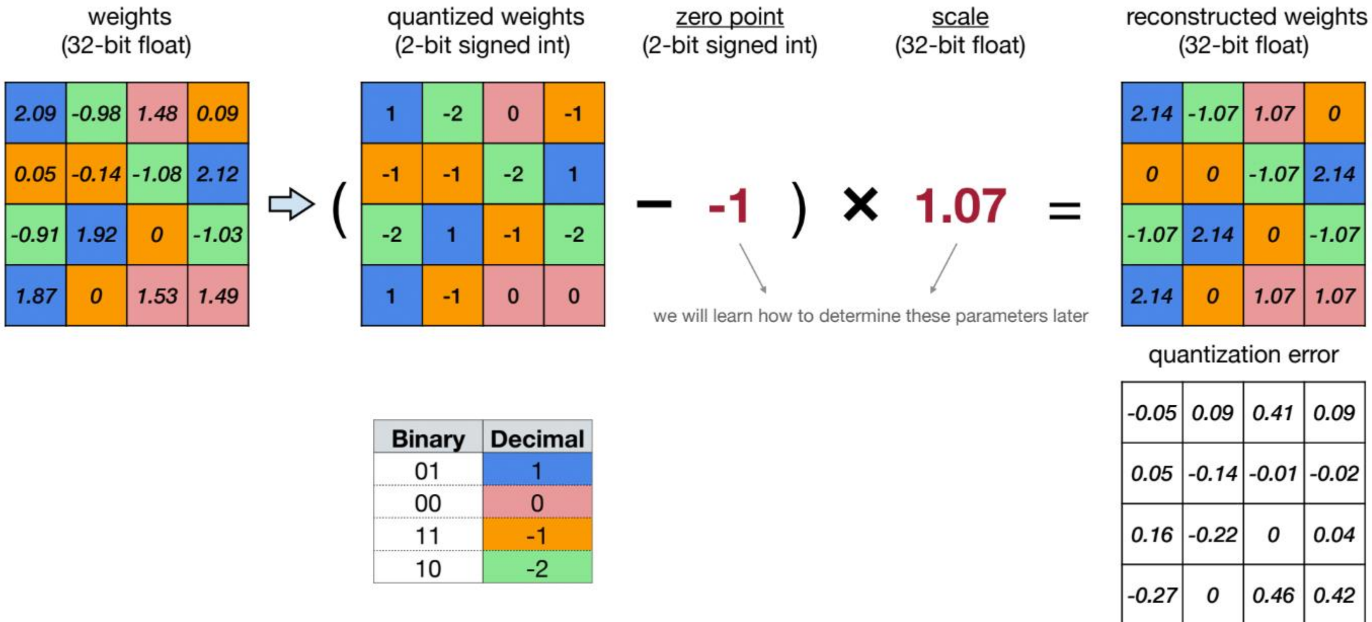
- Accuracy vs. compression rate for AlexNet on ImageNet dataset



Deep Compression [Han *et al.*, ICLR 2016]

What is Linear Quantization ?

- An affine mapping of integers to real numbers
- **Storage:** Integer Weights; **Computation:** Integer Arithmetic



weights
(32-bit float)

2.09	-0.98	1.48	0.09
0.05	-0.14	-1.08	2.12
-0.91	1.92	0	-1.03
1.87	0	1.53	1.49

⇨ (

quantized weights
(2-bit signed int)

1	-2	0	-1
-1	-1	-2	1
-2	1	-1	-2
1	-1	0	0

− -1) × 1.07 =

zero point
(2-bit signed int)

scale
(32-bit float)

we will learn how to determine these parameters later

reconstructed weights
(32-bit float)

2.14	-1.07	1.07	0
0	0	-1.07	2.14
-1.07	2.14	0	-1.07
2.14	0	1.07	1.07

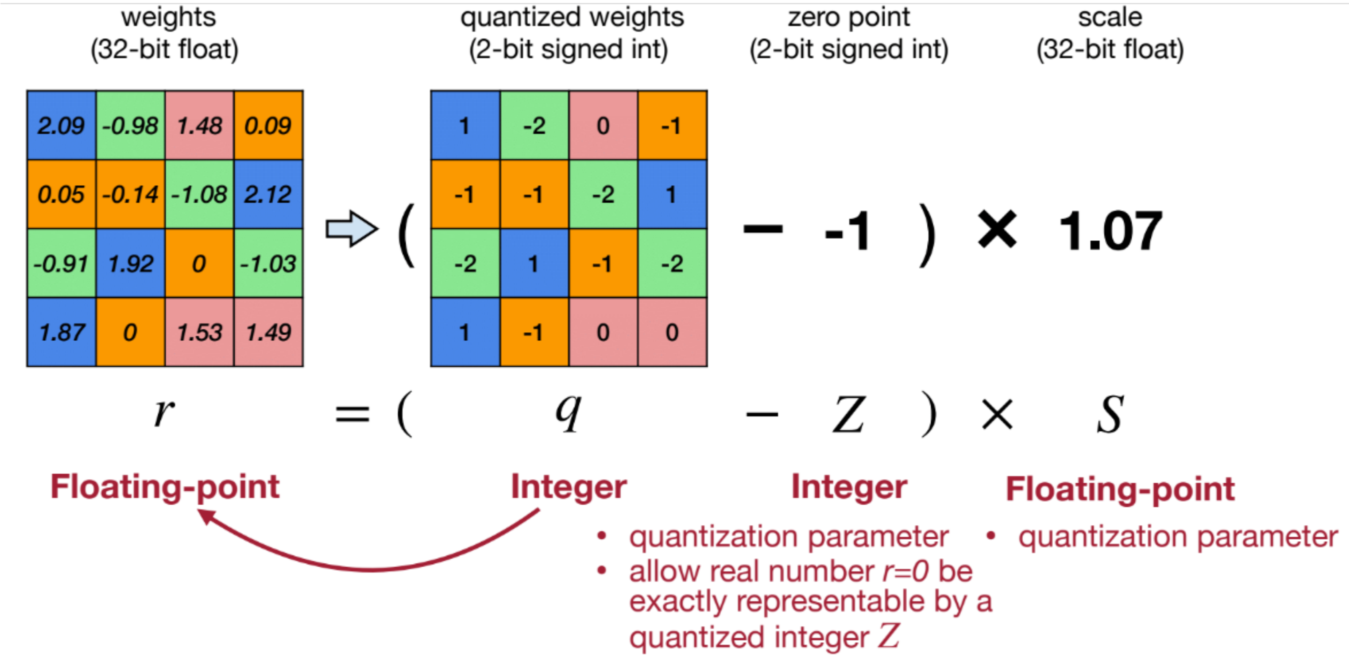
Binary	Decimal
01	1
00	0
11	-1
10	-2

quantization error

-0.05	0.09	0.41	0.09
0.05	-0.14	-0.01	-0.02
0.16	-0.22	0	0.04
-0.27	0	0.46	0.42

Linear Quantization

- An affine mapping of integers to real numbers (r = S(q - Z))



weights (32-bit float) | quantized weights (2-bit signed int) | zero point (2-bit signed int) | scale (32-bit float)

$$\begin{pmatrix} 2.09 & -0.98 & 1.48 & 0.09 \\ 0.05 & -0.14 & -1.08 & 2.12 \\ -0.91 & 1.92 & 0 & -1.03 \\ 1.87 & 0 & 1.53 & 1.49 \end{pmatrix} \Rightarrow \left(\begin{pmatrix} 1 & -2 & 0 & -1 \\ -1 & -1 & -2 & 1 \\ -2 & 1 & -1 & -2 \\ 1 & -1 & 0 & 0 \end{pmatrix} - -1 \right) \times 1.07$$

$$r = (q - Z) \times S$$

Floating-point | **Integer** | **Integer** | **Floating-point**

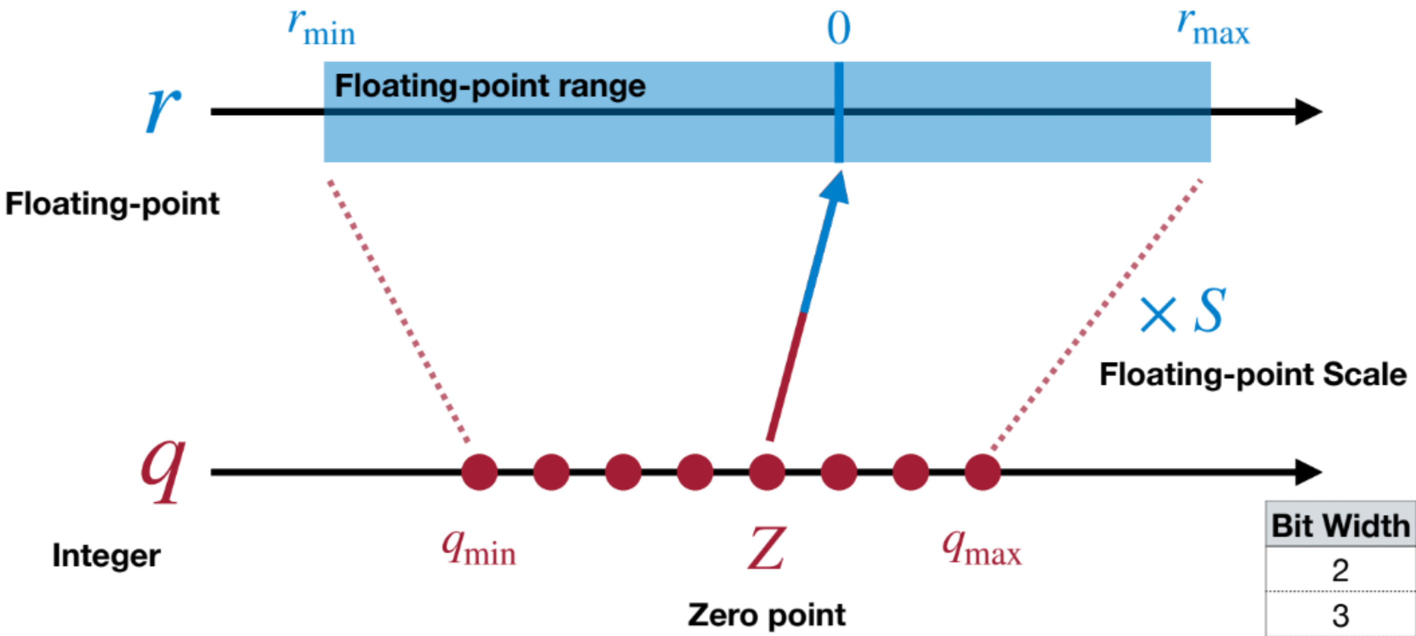
Integer:
- quantization parameter
- allow real number $r=0$ be exactly representable by a quantized integer Z

Floating-point:
- quantization parameter

Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference [Jacob *et al.*, CVPR 2018]

Linear Quantization

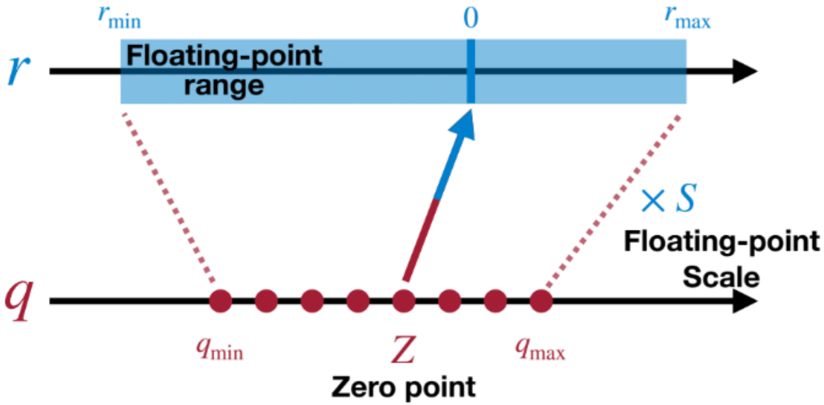
- An affine mapping of integers to real numbers (r = S(q - Z))



Bit Width	q_{min}	q_{max}
2	-2	1
3	-4	3
4	-8	7
N	-2^{N-1}	$2^{N-1}-1$

Scale of Linear Quantization

- An affine mapping of integers to real numbers (r = S(q - Z))



$$r_{max} = S\left(q_{max} - Z\right)$$

$$r_{min} = S\left(q_{min} - Z\right)$$

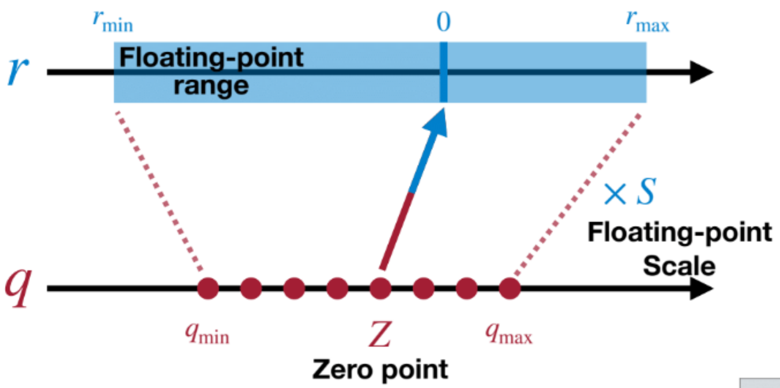
$$\downarrow$$

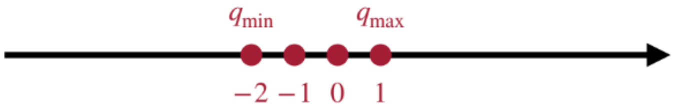
$$r_{max} - r_{min} = S\left(q_{max} - q_{min}\right)$$

$$S = \frac{r_{max} - r_{min}}{q_{max} - q_{min}}$$

Scale of Linear Quantization

- An affine mapping of integers to real numbers (r = S(q - Z))





2.09	-0.98	1.48	0.09
0.05	-0.14	-1.08	2.12
-0.91	1.92	0	-1.03
1.87	0	1.53	1.49

Binary	Decimal
01	1
00	0
11	-1
10	-2

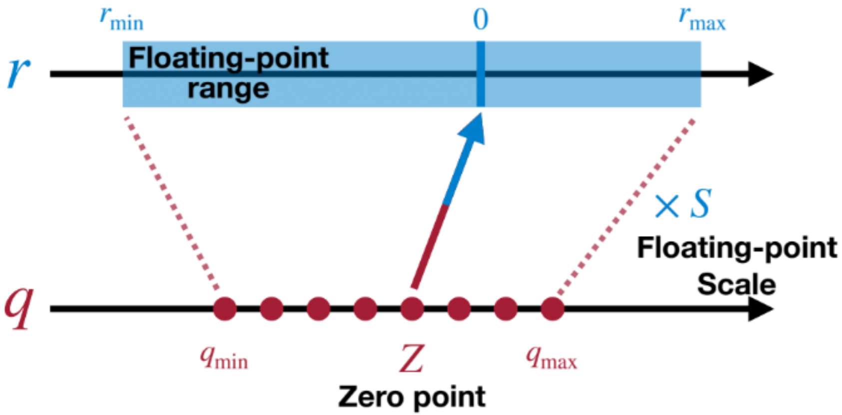
$$S = \frac{r_{\max} - r_{\min}}{q_{\max} - q_{\min}}$$

$$= \frac{2.12 - (-1.08)}{1 - (-2)}$$

$$= 1.07$$

Zero Point of Linear Quantization

- An affine mapping of integers to real numbers (r = S(q - Z))



$$r_{\min} = S\left(q_{\min} - Z\right)$$

$$\downarrow$$

$$Z = q_{\min} - \frac{r_{\min}}{S}$$

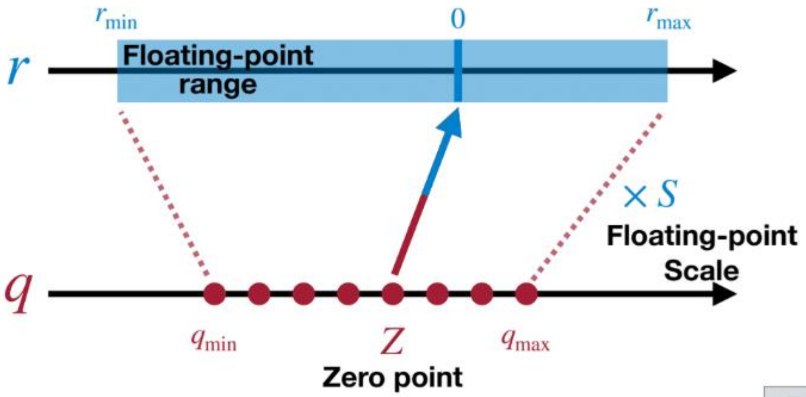
$$\downarrow$$

$$Z = \text{round}\left(q_{\min} - \frac{r_{\min}}{S}\right)$$

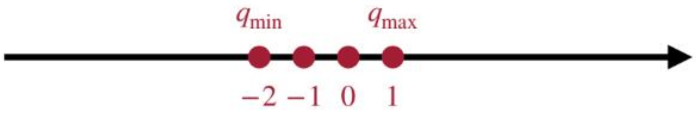
Zero Point of Linear Quantization

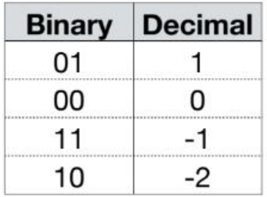
- An affine mapping of integers to real numbers (r = S(q - Z))

2.09	-0.98	1.48	0.09
0.05	-0.14	-1.08	2.12
-0.91	1.92	0	-1.03
1.87	0	1.53	1.49



$$Z = q_{\min} - \frac{r_{\min}}{S}$$





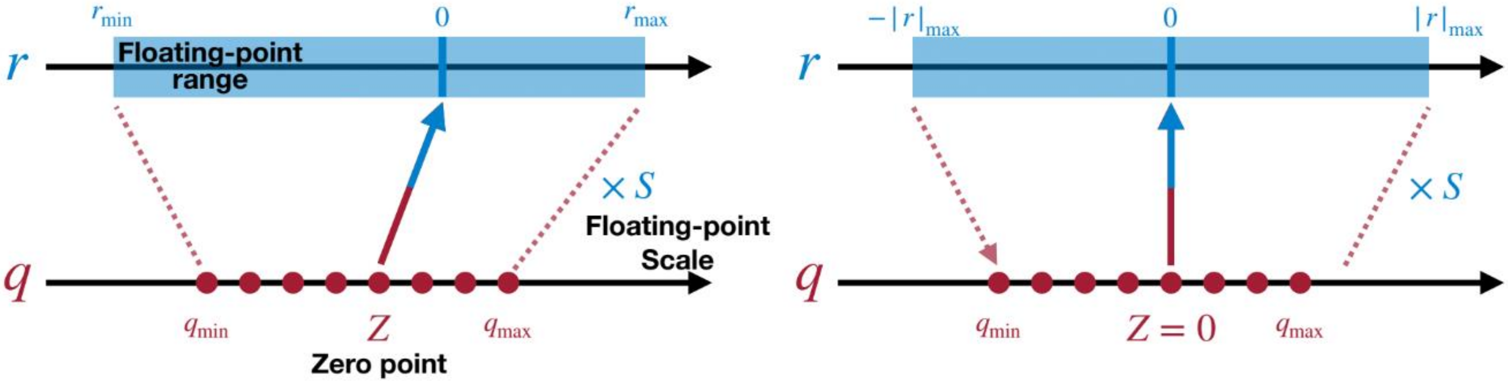
Binary	Decimal
01	1
00	0
11	-1
10	-2

$$= \text{round}(-2 - \frac{-1.08}{1.07})$$

$$= -1$$

Symmetric Linear Quantization

- Full range mode



$$S = \frac{r_{\max} - r_{\min}}{q_{\max} - q_{\min}}$$

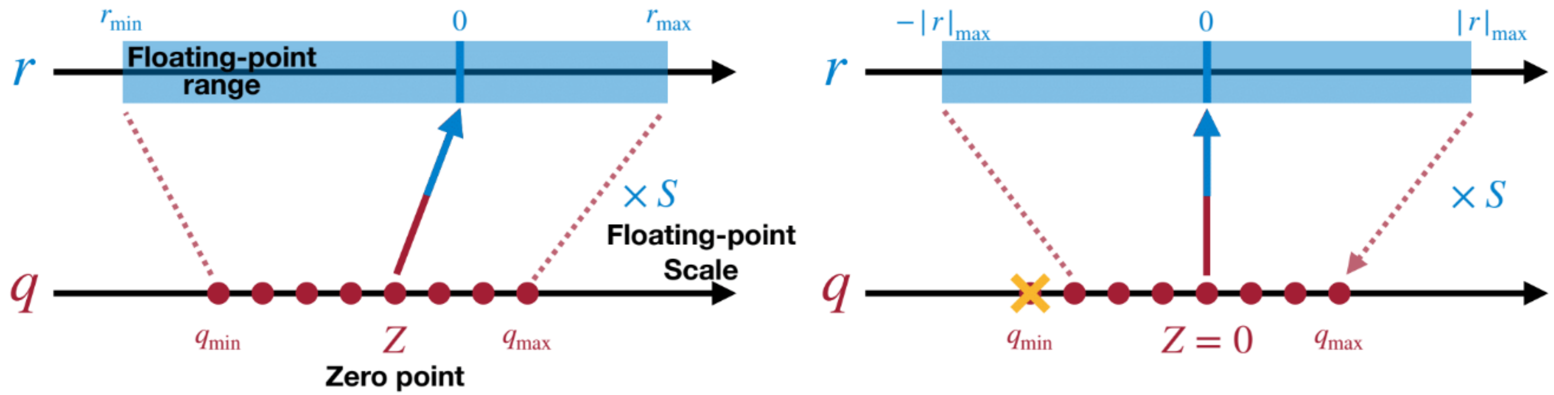
$$S = \frac{r_{\min}}{q_{\min} - Z} = \frac{-|r|_{\max}}{q_{\min}} = \frac{|r|_{\max}}{2^{N-1}}$$

- use full range of quantized integers
- example: PyTorch's native quantization, ONNX

Bit Width	$q_{\min}$	$q_{\max}$
2	-2	1
3	-4	3
4	-8	7
N	-2^{N-1}	$2^{N-1}-1$

Symmetric Linear Quantization

- Restricted range mode



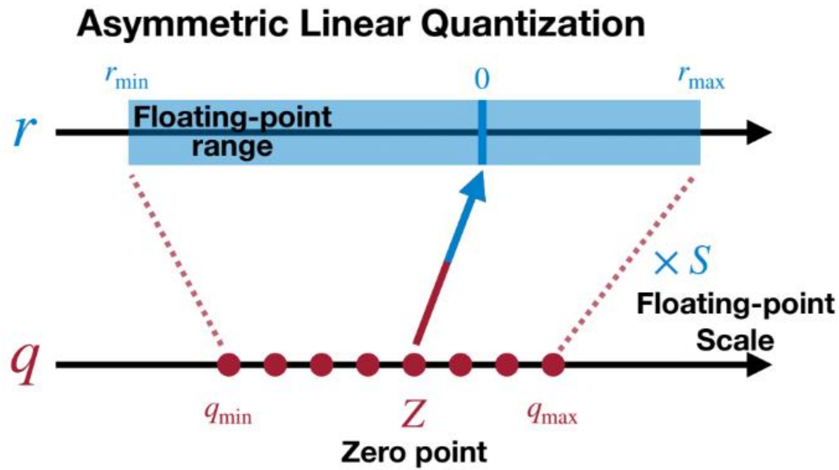
$$S = \frac{r_{\max} - r_{\min}}{q_{\max} - q_{\min}}$$

$$S = \frac{r_{\max}}{q_{\max} - Z} = \frac{|r|_{\max}}{q_{\max}} = \frac{|r|_{\max}}{2^{N-1} - 1}$$

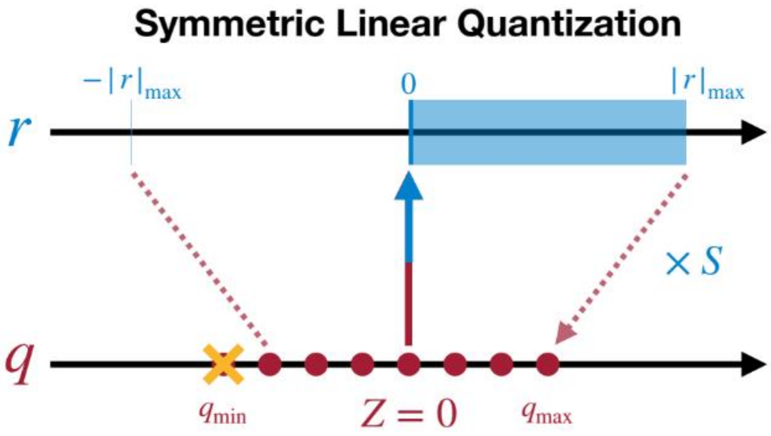
Bit Width	$q_{\min}$	$q_{\max}$
2	-2	1
3	-4	3
4	-8	7
N	-2^{N-1}	$2^{N-1}-1$

- example: TensorFlow, NVIDIA TensorRT, Intel DNNL

Asymmetric vs. Symmetric



- The quantized range is fully used.
- The implementation is more complex, and zero points require additional logic in hardware.



- The quantized range will be wasted for biased float range.
  - Activation tensor is non-negative after ReLU, and thus symmetric quantization will lose 1 bit effectively.
- The implementation is much simpler.

Neural Network Distiller

Linear Quantized Matrix Multiplication

- An affine mapping of integers to real numbers (r = S(q - Z))

$$\mathbf{Y} = \mathbf{W}\mathbf{X}$$

$$S_{\mathbf{Y}}\left(\mathbf{q}_{\mathbf{Y}} - Z_{\mathbf{Y}}\right) = S_{\mathbf{W}}\left(\mathbf{q}_{\mathbf{W}} - Z_{\mathbf{W}}\right) \cdot S_{\mathbf{X}}\left(\mathbf{q}_{\mathbf{X}} - Z_{\mathbf{X}}\right)$$

$$\mathbf{q}_{\mathbf{Y}} = \frac{S_{\mathbf{W}}S_{\mathbf{X}}}{S_{\mathbf{Y}}}\left(\mathbf{q}_{\mathbf{W}} - Z_{\mathbf{W}}\right)\left(\mathbf{q}_{\mathbf{X}} - Z_{\mathbf{X}}\right) + Z_{\mathbf{Y}}$$

$$\mathbf{q}_{\mathbf{Y}} = \frac{S_{\mathbf{W}}S_{\mathbf{X}}}{S_{\mathbf{Y}}}\left(\mathbf{q}_{\mathbf{W}}\mathbf{q}_{\mathbf{X}} - Z_{\mathbf{W}}\mathbf{q}_{\mathbf{X}} - Z_{\mathbf{X}}\mathbf{q}_{\mathbf{W}} + Z_{\mathbf{W}}Z_{\mathbf{X}}\right) + Z_{\mathbf{Y}}$$

Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference [Jacob *et al.*, CVPR 2018]

Linear Quantized Matrix Multiplication

- An affine mapping of integers to real numbers (r = S(q - Z))
    - Consider the following matrix multiplication

$$\mathbf{Y} = \mathbf{W}\mathbf{X}$$

$$\mathbf{q}_{\mathbf{Y}} = \boxed{\frac{S_{\mathbf{W}} S_{\mathbf{X}}}{S_{\mathbf{Y}}}} \left(\boxed{\mathbf{q}_{\mathbf{W}} \mathbf{q}_{\mathbf{X}} - Z_{\mathbf{W}} \mathbf{q}_{\mathbf{X}} \overbrace{- Z_{\mathbf{X}} \mathbf{q}_{\mathbf{W}} + Z_{\mathbf{W}} Z_{\mathbf{X}}}^{\text{Precompute}}} \right) + \boxed{Z_{\mathbf{Y}}}$$

Rescale to
***N*-bit Integer**

***N*-bit Integer Multiplication**
32-bit Integer Addition/Subtraction

***N*-bit Integer**
Addition

Linear Quantized Matrix Multiplication

- Consider the following matrix multiplication.

$$\mathbf{Y} = \mathbf{W}\mathbf{X}$$

$$\mathbf{q}_{\mathbf{Y}} = \boxed{\frac{S_{\mathbf{W}} S_{\mathbf{X}}}{S_{\mathbf{Y}}}} \left(\boxed{\mathbf{q}_{\mathbf{W}} \mathbf{q}_{\mathbf{X}} - Z_{\mathbf{W}} \mathbf{q}_{\mathbf{X}} - Z_{\mathbf{X}} \mathbf{q}_{\mathbf{W}} + Z_{\mathbf{W}} Z_{\mathbf{X}}} \right) + \boxed{Z_{\mathbf{Y}}}$$

- Empirically, the scale $\frac{S_{\mathbf{W}} S_{\mathbf{X}}}{S_{\mathbf{Y}}}$ is always in the interval (0, 1).

Fixed-point Multiplication

$$\frac{S_{\mathbf{W}} S_{\mathbf{X}}}{S_{\mathbf{Y}}} = 2^{-n} M_0, \quad \text{where} \quad M_0 \in [0.5, 1)$$

Bit Shift

Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference [Jacob *et al.*, CVPR 2018]

Linear Quantized Matrix Multiplication

$$\mathbf{Y} = \mathbf{W}\mathbf{X}$$

Precompute

$$\mathbf{q_Y} = \boxed{\frac{S_{\mathbf{W}} S_{\mathbf{X}}}{S_{\mathbf{Y}}}} \left(\boxed{\mathbf{q_W}\mathbf{q_X} - Z_{\mathbf{W}}\mathbf{q_X} - Z_{\mathbf{X}}\mathbf{q_W} + Z_{\mathbf{W}} Z_{\mathbf{X}}} \right) + \boxed{Z_{\mathbf{Y}}}$$

Rescale to *N*-bit Integer

***N*-bit Integer Multiplication**
32-bit Integer Addition/Subtraction

***N*-bit Integer Addition**

$$\mathbf{q_Y} = \boxed{\frac{S_{\mathbf{W}} S_{\mathbf{X}}}{S_{\mathbf{Y}}}} \left(\boxed{\mathbf{q_W}\mathbf{q_X} - Z_{\mathbf{X}}\mathbf{q_W}} \right) + \boxed{Z_{\mathbf{Y}}}$$

$$Z_{\mathbf{W}} = 0$$

Linear Quantized Fully-Connected Layer

- An affine mapping of integers to real numbers (r = S(q - Z))
    - Now, we consider the following fully-connected layer with bias

$$\mathbf{Y} = \mathbf{W}\mathbf{X} + \mathbf{b}$$

$$S_{\mathbf{Y}}\left(\mathbf{q}_{\mathbf{Y}} - Z_{\mathbf{Y}}\right) = S_{\mathbf{W}}\left(\mathbf{q}_{\mathbf{W}} - Z_{\mathbf{W}}\right) \cdot S_{\mathbf{X}}\left(\mathbf{q}_{\mathbf{X}} - Z_{\mathbf{X}}\right) + S_{\mathbf{b}}\left(\mathbf{q}_{\mathbf{b}} - Z_{\mathbf{b}}\right)$$

$$\downarrow \; Z_{\mathbf{W}} = 0$$

$$S_{\mathbf{Y}}\left(\mathbf{q}_{\mathbf{Y}} - Z_{\mathbf{Y}}\right) = \underline{\boxed{S_{\mathbf{W}} S_{\mathbf{X}}}\left(\mathbf{q}_{\mathbf{W}}\mathbf{q}_{\mathbf{X}} - Z_{\mathbf{X}}\mathbf{q}_{\mathbf{W}}\right)} + \underline{\boxed{S_{\mathbf{b}}}\left(\mathbf{q}_{\mathbf{b}} - \boxed{Z_{\mathbf{b}}}\right)}$$

Linear Quantized Fully-Connected Layer

- An affine mapping of integers to real numbers (r = S(q - Z))
    - Now, we consider the following fully-connected layer with bias

$$\mathbf{Y} = \mathbf{W}\mathbf{X} + \mathbf{b}$$

$$S_{\mathbf{Y}}\left(\mathbf{q}_{\mathbf{Y}} - Z_{\mathbf{Y}}\right) = S_{\mathbf{W}}\left(\mathbf{q}_{\mathbf{W}} - Z_{\mathbf{W}}\right) \cdot S_{\mathbf{X}}\left(\mathbf{q}_{\mathbf{X}} - Z_{\mathbf{X}}\right) + S_{\mathbf{b}}\left(\mathbf{q}_{\mathbf{b}} - Z_{\mathbf{b}}\right)$$

$$\downarrow \; Z_{\mathbf{W}} = 0$$

$$S_{\mathbf{Y}}\left(\mathbf{q}_{\mathbf{Y}} - Z_{\mathbf{Y}}\right) = S_{\mathbf{W}}S_{\mathbf{X}}\left(\mathbf{q}_{\mathbf{W}}\mathbf{q}_{\mathbf{X}} - Z_{\mathbf{X}}\mathbf{q}_{\mathbf{W}}\right) + S_{\mathbf{b}}\left(\mathbf{q}_{\mathbf{b}} - Z_{\mathbf{b}}\right)$$

$$\downarrow \; Z_{\mathbf{b}} = 0, \quad S_{\mathbf{b}} = S_{\mathbf{W}}S_{\mathbf{X}}$$

$$S_{\mathbf{Y}}\left(\mathbf{q}_{\mathbf{Y}} - Z_{\mathbf{Y}}\right) = S_{\mathbf{W}}S_{\mathbf{X}}\left(\mathbf{q}_{\mathbf{W}}\mathbf{q}_{\mathbf{X}} - Z_{\mathbf{X}}\mathbf{q}_{\mathbf{W}} + \mathbf{q}_{\mathbf{b}}\right)$$

Linear Quantized Fully-Connected Layer

- An affine mapping of integers to real numbers (r = S(q - Z))
  - Now, we consider the following fully-connected layer with bias

$$\mathbf{Y} = \mathbf{W}\mathbf{X} + \mathbf{b}$$

$$Z_{\mathbf{W}} = 0 \;\downarrow\; Z_{\mathbf{b}} = 0, \quad S_{\mathbf{b}} = S_{\mathbf{W}} S_{\mathbf{X}}$$

$$S_{\mathbf{Y}}\left(\mathbf{q_Y} - Z_{\mathbf{Y}}\right) = S_{\mathbf{W}} S_{\mathbf{X}}\left(\mathbf{q_W}\mathbf{q_X} - Z_{\mathbf{X}}\mathbf{q_W} + \mathbf{q_b}\right)$$

$$\downarrow$$

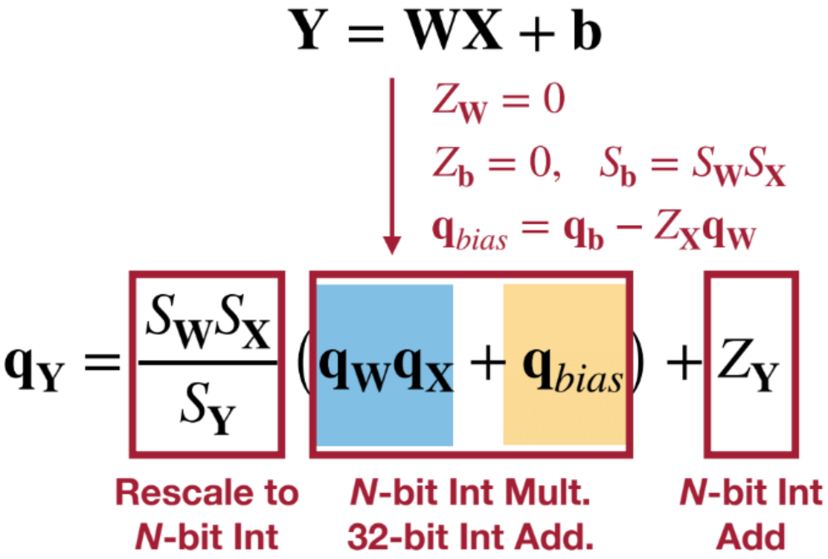
$$\mathbf{q_Y} = \frac{S_{\mathbf{W}} S_{\mathbf{X}}}{S_{\mathbf{Y}}}\left(\mathbf{q_W}\mathbf{q_X} + \overset{\text{Precompute}}{\mathbf{q_b} - Z_{\mathbf{X}}\mathbf{q_W}}\right) + Z_{\mathbf{Y}}$$

$$\downarrow \; \mathbf{q}_{bias} = \mathbf{q_b} - Z_{\mathbf{X}}\mathbf{q_W}$$

$$\mathbf{q_Y} = \frac{S_{\mathbf{W}} S_{\mathbf{X}}}{S_{\mathbf{Y}}}\left(\mathbf{q_W}\mathbf{q_X} + \mathbf{q}_{bias}\right) + Z_{\mathbf{Y}}$$

Linear Quantized Fully-Connected Layer

- An affine mapping of integers to real numbers (r = S(q - Z))
  - Now, we consider the following fully-connected layer with bias



$$\mathbf{Y} = \mathbf{W}\mathbf{X} + \mathbf{b}$$

$$Z_{\mathbf{W}} = 0$$

$$Z_{\mathbf{b}} = 0, \quad S_{\mathbf{b}} = S_{\mathbf{W}} S_{\mathbf{X}}$$

$$\mathbf{q}_{bias} = \mathbf{q}_{\mathbf{b}} - Z_{\mathbf{X}} \mathbf{q}_{\mathbf{W}}$$

$$\mathbf{q}_{\mathbf{Y}} = \boxed{\frac{S_{\mathbf{W}} S_{\mathbf{X}}}{S_{\mathbf{Y}}}} \left(\boxed{\mathbf{q}_{\mathbf{W}}\mathbf{q}_{\mathbf{X}} + \mathbf{q}_{bias}}\right) + \boxed{Z_{\mathbf{Y}}}$$

Rescale to *N*-bit Int

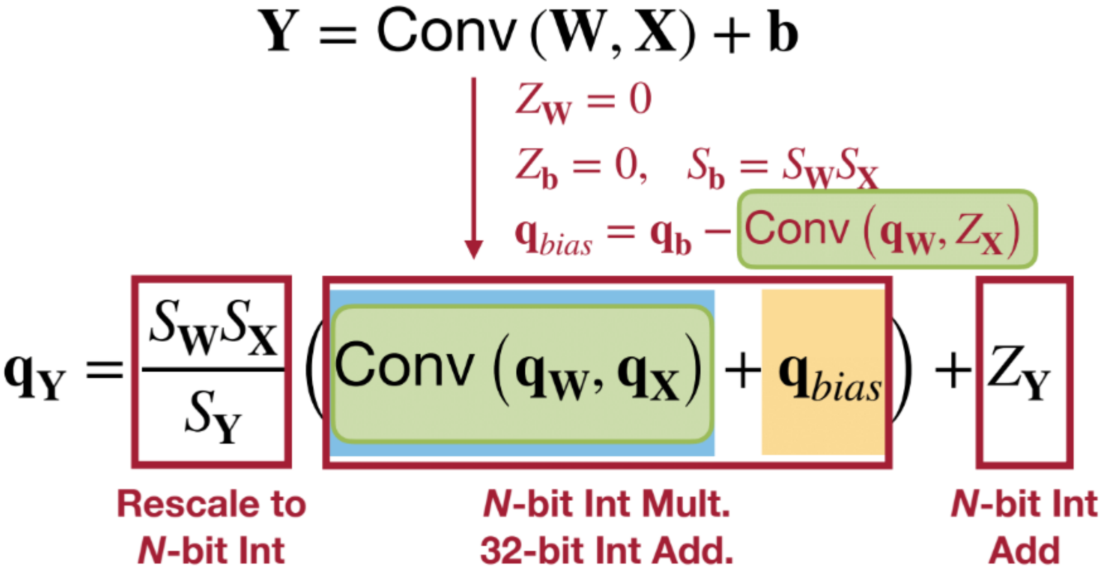
***N*-bit Int Mult. 32-bit Int Add.**

***N*-bit Int Add**

Note: both $\mathbf{q}_{\mathbf{b}}$ and $\mathbf{q}_{bias}$ are 32 bits.

Linear Quantized Convolution Layer

- An affine mapping of integers to real numbers (r = S(q - Z))
  - Now, we consider the following convolution layer



$$\mathbf{Y} = \text{Conv}\left(\mathbf{W}, \mathbf{X}\right) + \mathbf{b}$$

$$\downarrow \quad \begin{aligned} & Z_{\mathbf{W}} = 0 \\ & Z_{\mathbf{b}} = 0, \quad S_{\mathbf{b}} = S_{\mathbf{W}} S_{\mathbf{X}} \\ & \mathbf{q}_{bias} = \mathbf{q}_{\mathbf{b}} - \text{Conv}\left(\mathbf{q}_{\mathbf{W}}, Z_{\mathbf{X}}\right) \end{aligned}$$

$$\mathbf{q}_{\mathbf{Y}} = \boxed{\frac{S_{\mathbf{W}} S_{\mathbf{X}}}{S_{\mathbf{Y}}}} \left(\boxed{\text{Conv}\left(\mathbf{q}_{\mathbf{W}}, \mathbf{q}_{\mathbf{X}}\right) + \mathbf{q}_{bias}} \right) + \boxed{Z_{\mathbf{Y}}}$$

Rescale to *N*-bit Int

***N*-bit Int Mult.
32-bit Int Add.**

***N*-bit Int Add**

Note: both $\mathbf{q}_{\mathbf{b}}$ and $\mathbf{q}_{bias}$ are 32 bits.

Linear Quantized Convolution Layer

- An affine mapping of integers to real numbers (r = S(q - Z))
    - Now, we consider the following convolution layer

$$\mathbf{Y} = \text{Conv}(\mathbf{W}, \mathbf{X}) + \mathbf{b}$$

$$\Bigg\downarrow \quad \begin{array}{l} Z_{\mathbf{W}} = 0 \\ Z_{\mathbf{b}} = 0, \quad S_{\mathbf{b}} = S_{\mathbf{W}} S_{\mathbf{X}} \\ \mathbf{q}_{bias} = \mathbf{q}_{\mathbf{b}} - \boxed{\text{Conv}(\mathbf{q}_{\mathbf{W}}, Z_{\mathbf{X}})} \end{array}$$

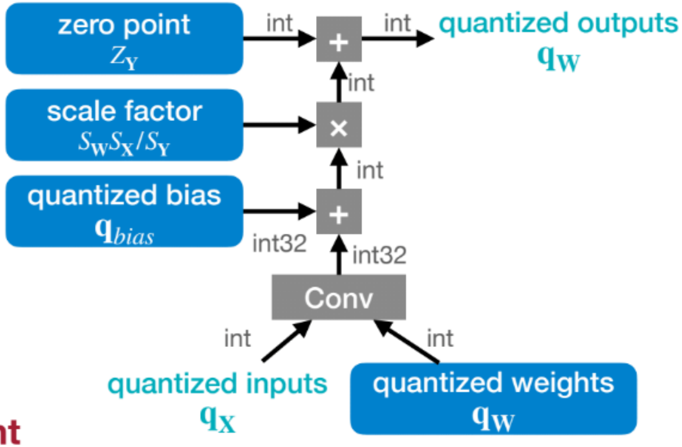
$$\mathbf{q}_{\mathbf{Y}} = \boxed{\frac{S_{\mathbf{W}} S_{\mathbf{X}}}{S_{\mathbf{Y}}}} \left(\boxed{\boxed{\text{Conv}(\mathbf{q}_{\mathbf{W}}, \mathbf{q}_{\mathbf{X}})} + \mathbf{q}_{bias}} \right) + \boxed{Z_{\mathbf{Y}}}$$

Rescale to *N*-bit Int

***N*-bit Int Mult.**
32-bit Int Add.

***N*-bit Int Add**

Note: both $\mathbf{q}_{\mathbf{b}}$ and $\mathbf{q}_{bias}$ are 32 bits.



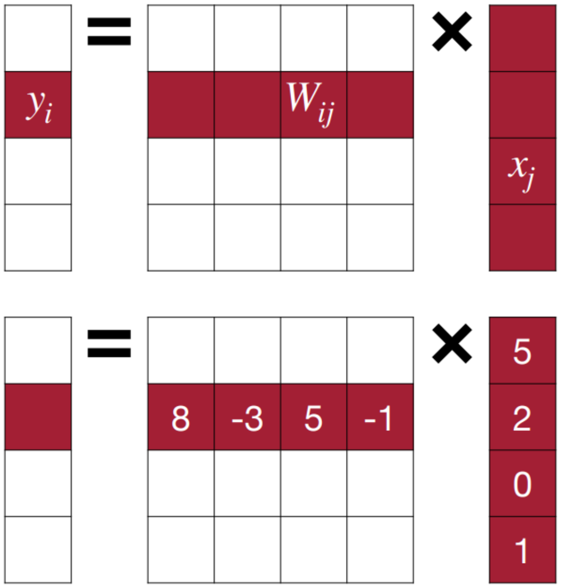
Binary/Ternary Quantization

- Could we push the quantization precision to 1 bit?

$$y_i = \sum_j W_{ij} \cdot x_j$$

$$= 8\times5 + (-3)\times2 + 5\times0 + (-1)\times1$$

input	weight	operations	memory	computation
$\mathbb{R}$	$\mathbb{R}$	+ ×	1×	1×

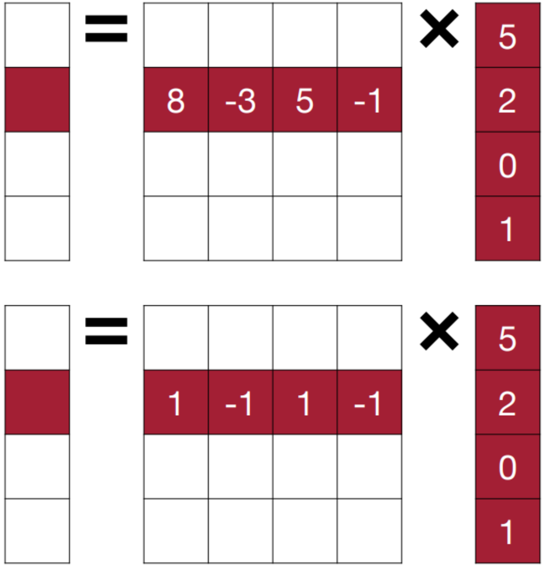


Binary/Ternary Quantization

- If weights are quantized to +1 and -1

$$y_i = \sum_j W_{ij} \cdot x_j$$

$$= 5 - 2 + 0 - 1$$



input	weight	operations	memory	computation
$\mathbb{R}$	$\mathbb{R}$	+ ×	1×	1×
$\mathbb{R}$	$\mathbb{B}$	+ -	~32× less	~2× less

BinaryConnect: Training Deep Neural Networks with Binary Weights during Propagations [Courbariaux *et al.*, NeurIPS 2015]
XNOR-Net: ImageNet Classification using Binary Convolutional Neural Networks [Rastegari *et al.*, ECCV 2016]

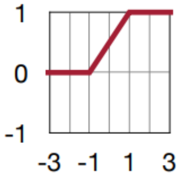
Binarization

- **Deterministic Binarization**
  - directly computes the bit value based on a threshold, usually 0, resulting in a sign function.

$$q = \text{sign}\,(r) = \begin{cases} +1, & r \geq 0 \\ -1, & r < 0 \end{cases}$$

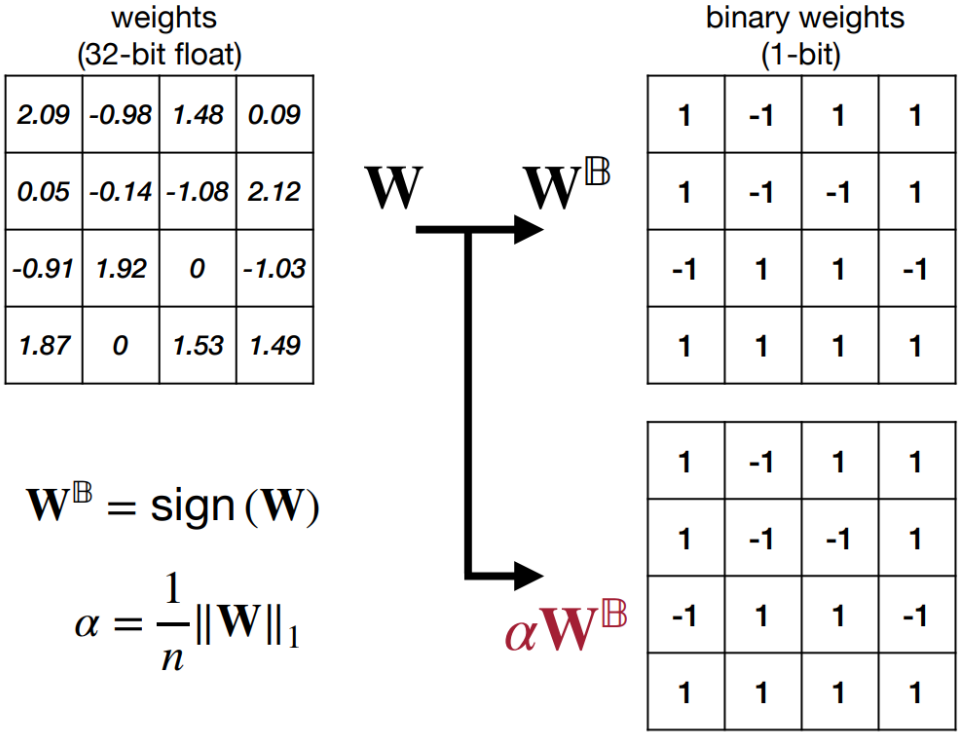
- **Stochastic Binarization**
  - use global statistics or the value of input data to determine the probability of being -1 or +1
    - *e.g.*, in Binary Connect (BC), probability is determined by hard sigmoid function $\sigma(r)$

$$q = \begin{cases} +1, & \text{with probability } p = \sigma(r) \\ -1, & \text{with probability } 1 - p \end{cases}, \quad \text{where } \sigma(r) = \min(\max(\frac{r+1}{2},0),1)$$



    - harder to implement as it requires the hardware to generate random bits when quantizing.

Minimizing Quantization Error in Binarization



weights
(32-bit float)

2.09	*-0.98*	*1.48*	*0.09*
0.05	*-0.14*	*-1.08*	*2.12*
-0.91	*1.92*	*0*	*-1.03*
1.87	*0*	*1.53*	*1.49*

$\mathbf{W}$

$\mathbf{W}^{\mathbb{B}}$

binary weights
(1-bit)

1	-1	1	1
1	-1	-1	1
-1	1	1	-1
1	1	1	1

$$\|\mathbf{W} - \mathbf{W}^{\mathbb{B}}\|_F^2 = 9.28$$

$$\mathbf{W}^{\mathbb{B}} = \text{sign}\,(\mathbf{W})$$

$$\alpha = \frac{1}{n}\|\mathbf{W}\|_1$$

$\alpha\mathbf{W}^{\mathbb{B}}$

1	-1	1	1
1	-1	-1	1
-1	1	1	-1
1	1	1	1

scale
(32-bit float)

$$\times\ \mathbf{1.05} = \frac{1}{16}\|\mathbf{W}\|_1$$

$$\|\mathbf{W} - \alpha\mathbf{W}^{\mathbb{B}}\|_F^2 = 9.24$$

AlexNet-based Network	ImageNet Top-1 Accuracy Delta
BinaryConnect	-21.2%
Binary Weight Network (BWN)	0.2%

XNOR-Net: ImageNet Classification using Binary Convolutional Neural Networks [Rastegari *et al.*, ECCV 2016]

Binary Net

- **Binary Connect**
    - Weights {-1, 1} (Bipolar binary), Activation 32-bit float
    - Accuracy loss: 19 % on AlexNet
- **Binarized Neural Networks**
    - Weights {-1, 1}, Activations {-1, 1}
    - Both of operands are binary, the multiplication turns into an XNOR
    - Accuracy loss: 29.8 % on AlexNet

XNOR

A	B	Out
0	0	1
1	0	0
0	1	0
1	1	1

Popcount (110010001) = 4

```
for each i in width:
    C += A[row][i] * B[i][col]
```

```
for each i in width:
    C += popcount(XNOR(A[row][i] * B[i][col]))
```

Courbariaux., NeurIPS, 2015

Case Study: Binary Multiplication

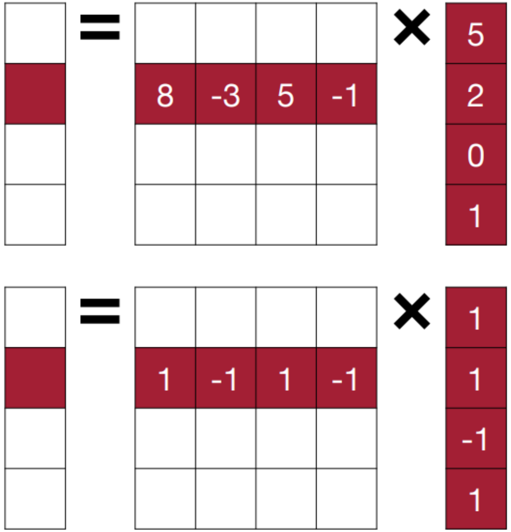
- A = 10010, B = 01111 (0 is really -1 here)
- **Dot product**:
  - A * B = (1 * -1) + (-1 * 1) + (-1 * 1) + (1 * 1) + (-1 * 1) = **-3**
- P = XNOR (A, B) = 00010, popcount(P) = 1
- Result = 2 * P – N, where N is the total number of bits
- 2 * P – N = 2 * 1 – 5 = **-3**

https://sushscience.wordpress.com/2017/10/01/understanding-binary-neural-networks/

XNOR-Net

- If both activations and weights are binarized

$$
\begin{aligned}
y_i &= \sum_j W_{ij} \cdot x_j \\
&= 1\times1 + (-1)\times1 + 1\times(-1) + (-1)\times1 \\
&= 1 + (-1) + (-1) + (-1) = -2
\end{aligned}
$$



XNOR-Net: ImageNet Classification using Binary Convolutional Neural Networks [Rastegari *et al.*, ECCV 2016]

XNOR-Net

- If both activations and weights are binarized

$$
\begin{aligned}
y_i &= \sum_j W_{ij} \cdot x_j \\
&= \boxed{1\times1 + (-1)\times1 + 1\times(-1) + (-1)\times1} \\
&= 1 + (-1) + (-1) + (-1) = -2
\end{aligned}
$$

W	X	Y=WX
1	1	1
1	-1	-1
-1	-1	1
-1	1	-1

b_W	b_X	XNOR(b_W, b_X)
1	1	1
1	0	0
0	0	1
0	1	0

XNOR-Net: ImageNet Classification using Binary Convolutional Neural Networks [Rastegari *et al.*, ECCV 2016]

XNOR-Net

- If both activations and weights are binarized

$$y_i = \sum_j W_{ij} \cdot x_j$$

$$= 1\times1 + (-1)\times1 + 1\times(-1) + (-1)\times1 \qquad = \boxed{1 \text{ xnor } 1 + 0 \text{ xnor } 1 + 1 \text{ xnor } 0 + 0 \text{ xnor } 1}$$

$$= 1 + (-1) + (-1) + (-1) = -2 \qquad = 1 + 0 + 0 + 0 \; = \; 1$$

?

W	X	Y=WX
1	1	1
1	-1	-1
-1	-1	1
-1	1	-1

b_W	b_X	XNOR(b_W, b_X)
1	1	1
1	0	0
0	0	1
0	1	0

XNOR-Net: ImageNet Classification using Binary Convolutional Neural Networks [Rastegari *et al.*, ECCV 2016]

XNOR-Net

- If both activations and weights are binarized

$$y_i = \sum_j W_{ij} \cdot x_j$$

$$= 1\times1 + (-1)\times1 + 1\times(-1) + (-1)\times1$$

$$= 1 + (-1) + (-1) + (-1) = -2$$

$$y_i = -n + 2 \cdot \sum_j W_{ij} \text{ xnor } x_j$$

$$= 1 \text{ xnor } 1 + 0 \text{ xnor } 1 + 1 \text{ xnor } 0 + 0 \text{ xnor } 1$$

$$= \underset{\uparrow +2}{1} + 0 + 0 + 0 = \boxed{\begin{array}{c} 1 \times 2 \\ + \\ -4 \end{array}} = -2$$

Assuming -1 -1 -1 -1 → -4

W	X	Y=WX
1	1	1
1	-1	-1
-1	-1	1
-1	1	-1

b_W	b_X	XNOR(b_W, b_X)
1	1	1
1	0	0
0	0	1
0	1	0

XNOR-Net: ImageNet Classification using Binary Convolutional Neural Networks [Rastegari *et al.*, ECCV 2016]

XNOR-Net

- If both activations and weights are binarized

$$y_i = -n + 2 \cdot \boxed{\sum_j} W_{ij} \text{ xnor } x_j \quad \rightarrow \quad y_i = -n + \text{popcount}\left(W_i \text{ xnor } x\right) \ll 1$$

$$= -4 + 2 \times (1 \text{ xnor } 1 + 0 \text{ xnor } 1 + 1 \text{ xnor } 0 + 0 \text{ xnor } 1)$$

$$= -4 + 2 \times \boxed{(1 + 0 + 0 + 0)} = -2$$

→ popcount: return the number of 1

W	X	Y=WX
1	1	1
1	-1	-1
-1	-1	1
-1	1	-1

b_W	b_X	XNOR(b_W, b_X)
1	1	1
1	0	0
0	0	1
0	1	0

XNOR-Net: ImageNet Classification using Binary Convolutional Neural Networks [Rastegari *et al.*, ECCV 2016]

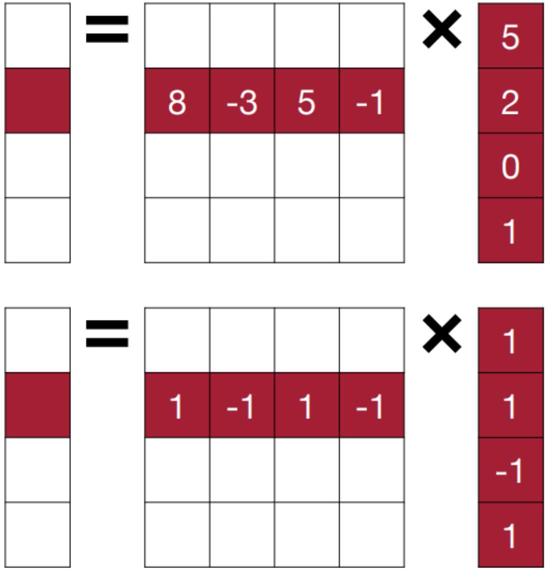
XNOR-Net

- If both activations and weights are binarized

$$y_i = -n + \text{popcount}\left(W_i \text{ xnor } x\right) \ll 1$$

$$= -4 + \text{popcount}(1010 \text{ xnor } 1101) \ll 1$$

$$= -4 + \text{popcount}(1000) \ll 1 = -4 + 2 = -2$$

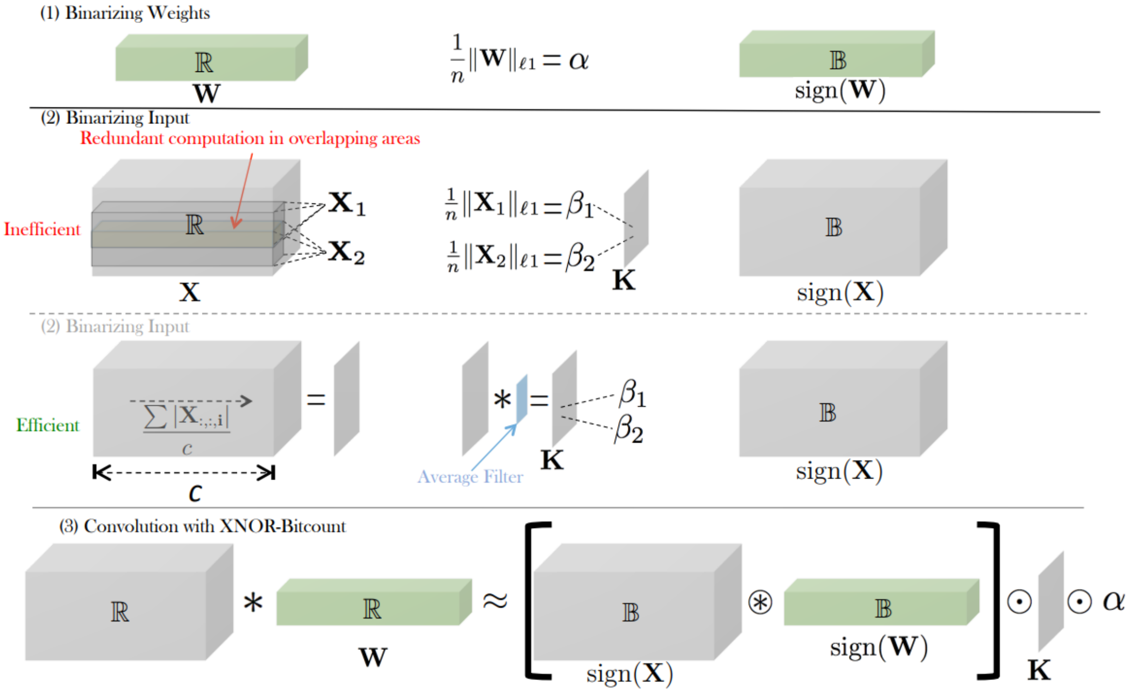


input	weight	operations	memory	computation
$\mathbb{R}$	$\mathbb{R}$	+ ×	1×	1×
$\mathbb{R}$	$\mathbb{B}$	+ -	~32× less	~2× less
$\mathbb{B}$	$\mathbb{B}$	xnor, popcount	~32× less	~58× less

XNOR-Net: ImageNet Classification using Binary Convolutional Neural Networks [Rastegari *et al.*, ECCV 2016]

XNOR-Net

- Minimizing quantization error in binarization



This slide is from XNOR-Net: ImageNet Classification using Binary Convolutional Neural Networks [Rastegari *et al.*, ECCV 2016]

XNOR-Net

Neural Network	Quantization	Bit-Width		ImageNet Top-1 Accuracy Delta
		W	A	
AlexNet	BWN	1	32	0.2%
	BNN	1	1	-28.7%
	XNOR-Net	1	1	-12.4%
GoogleNet	BWN	1	32	-5.80%
	BNN	1	1	-24.20%
ResNet-18	BWN	1	32	-8.5%
	XNOR-Net	1	1	-18.1%

* BWN: Binary Weight Network with scale for weight binarization
* BNN: Binarized Neural Network without scale factors
* XNOR-Net: scale factors for both activation and weight binarization

Binarized Neural Networks: Training Deep Neural Networks with Weights and Activations Constrained to +1 or −1. [Courbariaux *et al.*, Arxiv 2016]
XNOR-Net: ImageNet Classification using Binary Convolutional Neural Networks [Rastegari *et al.*, ECCV 2016]

Ternary Weight Networks (TWN)

Weights are quantized to +1, -1 and 0

$$q = \begin{cases} r_t, & r > \Delta \\ 0, & |r| \leq \Delta, \\ -r_t, & r < -\Delta \end{cases} \quad \text{where } \Delta = 0.7 \times \mathbb{E}\left(|r|\right), r_t = \mathbb{E}_{|r|>\Delta}\left(|r|\right)$$

weights $\mathbf{W}$
(32-bit float)

2.09	*-0.98*	*1.48*	*0.09*
0.05	*-0.14*	*-1.08*	*2.12*
-0.91	*1.92*	*0*	*-1.03*
1.87	*0*	*1.53*	*1.49*

→

ternary weights $\mathbf{W}^{\mathbb{T}}$
(2-bit)

1	*-1*	*1*	*0*
0	*0*	*-1*	*1*
-1	*1*	*0*	*-1*
1	*0*	*1*	*1*

$$\Delta = 0.7 \times \frac{1}{16}\|\mathbf{W}\|_1 = 0.73$$

$$\times \quad 1.5 = \frac{1}{11}\|\mathbf{W}_{\mathbf{W}^{\mathbb{T}}\neq 0}\|_1$$

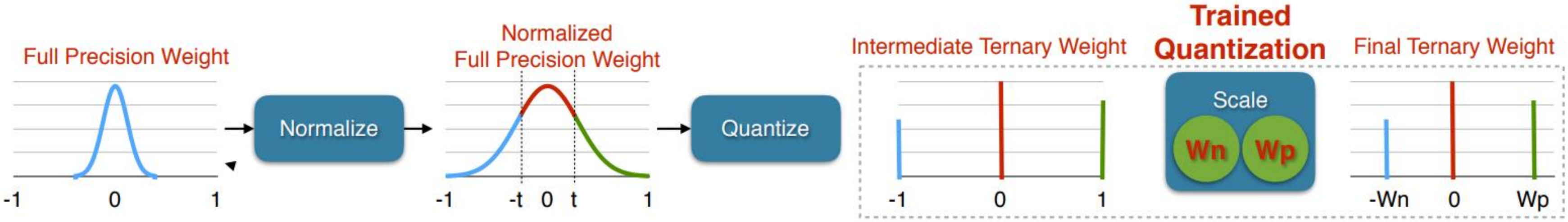
ImageNet Top-1 Accuracy	Full Precision	1 bit (BWN)	2 bit (TWN)
ResNet-18	69.6	60.8	65.3

Ternary Weight Networks [Li *et al.*, Arxiv 2016]

Ternary Weight Networks (TWN)

- Instead of using fixed scale r_t, TTQ introduces two *trainable* parameters w_p and w_n to represent the positive and negative scales in the quantization.

$$q = \begin{cases} w_p, & r > \Delta \\ 0, & |r| \le \Delta \\ -w_n, & r < -\Delta \end{cases}$$



ImageNet Top-1 Accuracy	Full Precision	1 bit (BWN)	2 bit (TWN)	TTQ
ResNet-18	69.6	60.8	65.3	66.6

Trained Ternary Quantization [Zhu *et al.*, ICLR 2017]

What do we Learn from Quantization?

- **Quantization** can improve DNN computational throughput while maintaining accuracy
- Layers on DNN models can be offered with **different bit widths**
- Varying bit width requires **the support of the hardware**
- **No systematic approach** to figure out the proper bit width in layers of DNN models
- What else ?

Takeaway Questions

- What are purposes of data quantization ?
    - (A) Constrain the value of inputs to a set of discrete values
    - (B) Create more values
    - (C) Improve the degree of parallelism on DNN training
- Why training requires large bit width ?
    - (A) The training needs to compute more data
    - (B) Avoid the value underflow and overflow
    - (C) Gradient and weight update have a larger range