

TinyML and Efficient Deep Learning Computing

Deep Learning Basics

From efficiency viewpoint, a wide and shallow network is good, because we can execute operation parallelly and make use of full compute available. From accuracy viewpoint network should be deep.

Full Connected layer: $y_i = \sum_j w_{ij} x_j + b_i$

Input feature X shape	(n, c _i)
Output feature Y shape	(n, c _o)
Weight W	(c _o , c _i)
Bias b	(c _o ,)

Convolution Layer:

	1D conv	2D conv	Grouped
Input feature X shape	(n, c _i , w _i)	(n, c _i , h _i , w _i)	(n, c _i , h _i , w _i)
Output feature Y shape	(n, c _o , w _o)	(n, c _o , h _o , w _o)	(n, c _o , h _o , w _o)
Weight W	(c _o , c _i , k _w)	(c _o , c _i , k _h , k _w)	(g.c _o /g, c _i /g, k _h , k _w)
Bias b	(c _o ,)	(c _o ,)	(c _o ,)

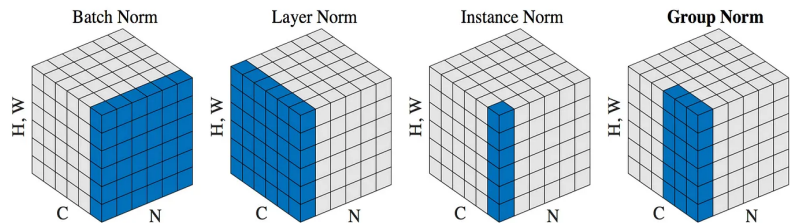
Depthwise Convolution Layer: Independent filter for each channel $g = c_i = c_o$

Normalization Layer

$$\hat{x}_k = \frac{1}{\sigma_i} (x_k - \mu_i)$$

$$\mu_i = \frac{1}{m} \sum_{k \in S_i} x_k$$

$$\sigma_i = \sqrt{\frac{1}{m} \sum_{k \in S_i} (x_k - \mu_i)^2}$$



Then learns a per-channel linear

transform (trainable scale γ and shift β) to compensate for possible lost of representational ability.

$$y = \gamma_{ic} \hat{x}_i + \beta_{ic}$$

Activation Function: Sigmoid, ReLU, ReLU6, Leaky ReLU, Swish, Hard Swish, Tanh, GELU, ELU

Efficiency Metrics

Memory Related: #parameters, model size, total/peak #activations

Computation-Related: MAC, FLOP, FLOPS, OP, OPS

Latency and throughput are not related to each other. Suppose an algo take 50ms to process an image, while other algo process 4 images parallelly in 100ms. Throughput of first algo is 20images/s while second is 40 images/s

Latency

$$\text{Latency} \approx \max(T_{\text{computation}}, T_{\text{memory}})$$

$$T_{\text{computation}} \approx \frac{\text{No. of operations in Neural Network Model}}{\text{No. of operation that processor can process per second}}$$

NN Specification

HW Specification

$$T_{\text{memory}} \approx T_{\text{data movement of activations}} + T_{\text{data movements of weight}}$$

$$T_{\text{data movement of activations}} \approx \frac{\text{Neural Network Model Size}}{\text{Memory BW of processor}}$$

NN Specification

HW Specification

$$T_{\text{data movements of weight}} \approx \frac{\text{Input activation size} + \text{Output Activation size}}{\text{Memory BW of processor}}$$

NN Specification

HW Specification

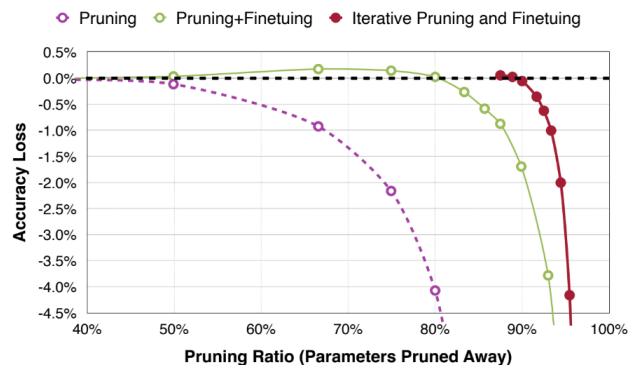
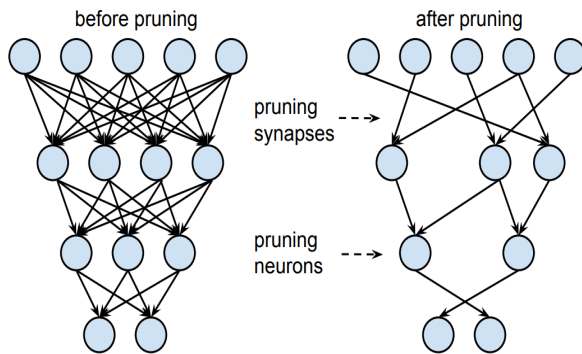
Data movement → More memory reference → More Energy

MAC Multiply-Accumulate operation (MAC) $a \leftarrow a + b * c$

Layer	MACs (Batch size n =1)
Linear Layer	$c_o * c_i$
Convolution	$c_i * k_h * k_w * h_o * w_o * c_o$
Grouped Convolution	$c_i / g * k_h * k_w * h_o * w_o * c_o$
Depthwise Convolution	$k_h * k_w * h_o * w_o * c_o$

One Multiply-accumulate operation is two floating point operations(FLOP)
Floating Point Operations Per Second (FLOPS)

Pruning



In general, we formulate as follows:

$$\arg \min_{W_p} L(x; W_p) \text{ subject to } \|W_p\|_0 < N$$

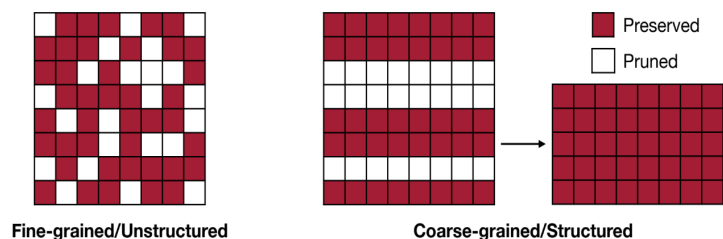
L represents the objective function of neural network, x is input, W is original weights, W_p is pruned weights. N is the target.

Fine -grained/Unstructured

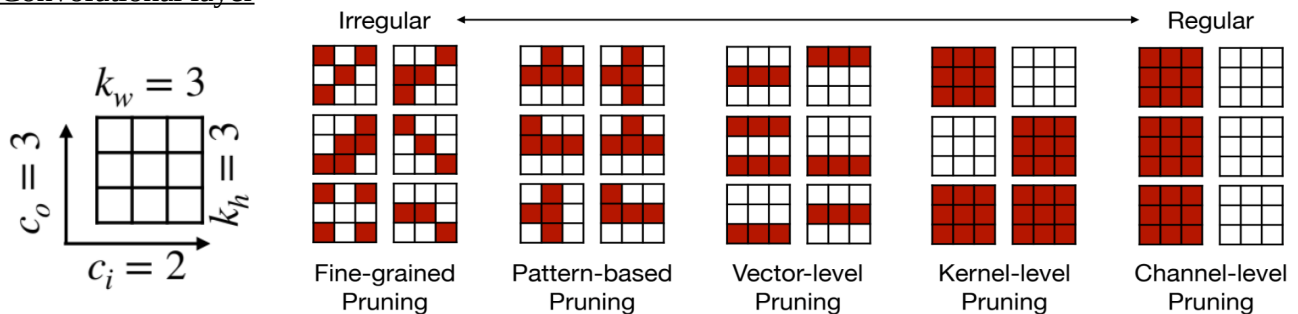
- More flexible pruning index choice
- Hard to accelerate (irregular)

Coarse-grained/Structured

- Less flexible pruning index choice
- Easy to accelerate



Convolutional layer

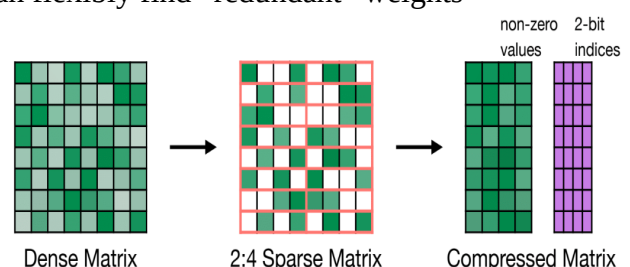


Fine grained Pruning

- Flexible Pruning Indices
- Usually larger compression ratio since we can flexibly find “redundant” weights
- Can deliver speed up on some custom hardware (EIE) but not GPU (easily)

Pattern-based Pruning N:M sparsity

- N:M sparsity means that in each contiguous M elements, N of them are pruned
- Classic case is 2:4



- Supported by NVIDIA's Ampere GPU architecture, which deliver 2x speed up.

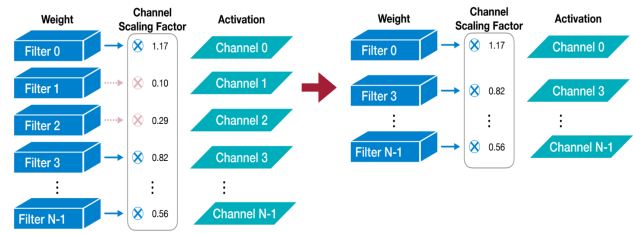
Channel Pruning

- Pro: Direct speed up due to reduced channel numbers (leading to an NN with smaller #channels)
- Con: Smaller compression ratio

Magnitude based Pruning: considers weights with larger absolute values are more important than other weights.

- For element-wise pruning, Importance = $|W|$
- For row-wise pruning, L_p -norm magnitude can be defined Importance = $\|W^{(s)}\|_p = \left(\sum_{i \in S} |w_i|^p\right)^{1/p}$

Scaling based Pruning: This pruning methods use scaling factors along with weight magnitudes for pruning. In this method, the trainable scaling factor is multiplied by the output of the channels. These scaling factors are used to calculate the importance of the channels. The filters/ output channels with a small scaling factor magnitude are pruned. The scaling factor can be reused from batch normalization layer



$$z_o = \gamma \frac{z_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} + \beta$$

Second-Order-based Pruning: Minimize the error on loss function introduced by pruning synapses. The induced error can be approximated by Taylor series

$$\delta L = L(x; W) - L(x; W_p = W - \delta W) = \sum_i g_i \delta w_i + \frac{1}{2} \sum_i h_{ii} \delta w_i^2 + \frac{1}{2} \sum_{i \neq j} h_{ij} \delta w_i \delta w_j + O(\|\delta W\|^3)$$

where, $g_i = \frac{\delta L}{\delta w_i}$ $h_{ij} = \frac{\delta^2 L}{\delta w_i \delta w_j}$

Optical Brain damage assume:

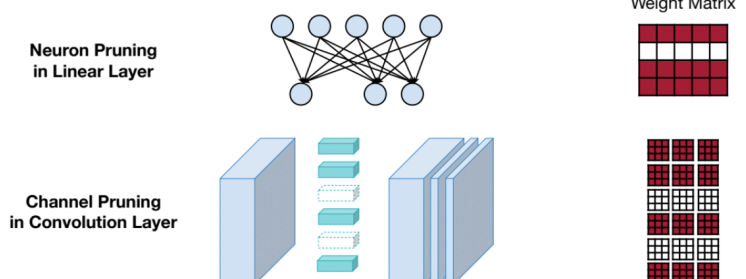
- L is nearly quadratic: last term is neglected
- NN Trainig has converged: first order terms are neglected
- Error caused by deleting each param is independent: cross terms are neglected.

$$\delta L_i = L(x; W) - L(x; W_p | w_i = 0) \approx \frac{1}{2} \sum_i h_{ii} \delta w_i^2$$

Synapses with smaller induced error $|\delta L_i|$ will be removed, Importance_{w_i} = $|\delta L_i| = \frac{1}{2} h_{ii} w_i^2$

Selection of Neurons to Prune:

Neuron pruning is coarse-grained weight pruning



Percentage of Zero-Based Pruning: ReLU activation will generate zeros in output activation. Similar to magnitude of weight, the Average Percentage of Zero activation (APoZ) can be exploited to measure the importance of neuron

Regression-based Pruning: Instead of considering the pruning error of the objective function, regression-based pruning minimizes the reconstruction error of the corresponding layer's outputs.

$$Z = XW^T = \sum_{c=0}^{c_i-1} X_c W_c^T$$

The problem can be formulated as

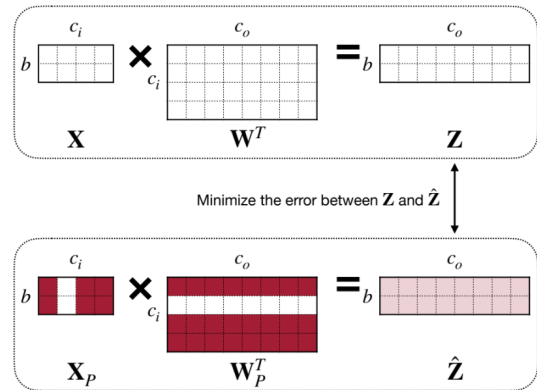
$$\arg \min_{W, \beta} \|Z - \hat{Z}\|_F^2 = \|Z - \sum_{c=0}^{c_i-1} \beta_c X_c W_c^T\|_F^2$$

subject to $\|\beta\|_0 \leq N_c$

$\beta_c = 0$ means channel c is pruned. N_c is number of nonzero channels.

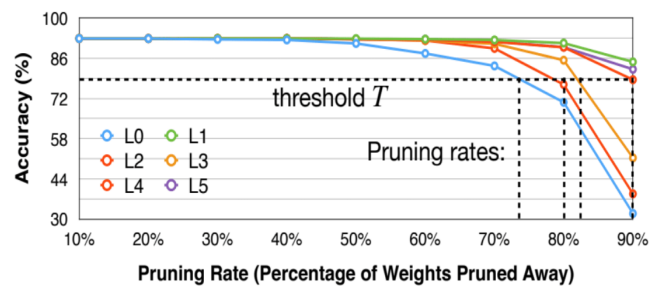
Solve:

- Fix W , solve β for selection
- Fix β , solve W to minimize reconstruction error



Pruning Ratio Non-uniform pruning is better than uniform shrinking

Analyse the sensitivity of each layer. We need different pruning ratio for each layer since different layer have different sensitivity. Some layers are more sensitive. Some layers are more redundant



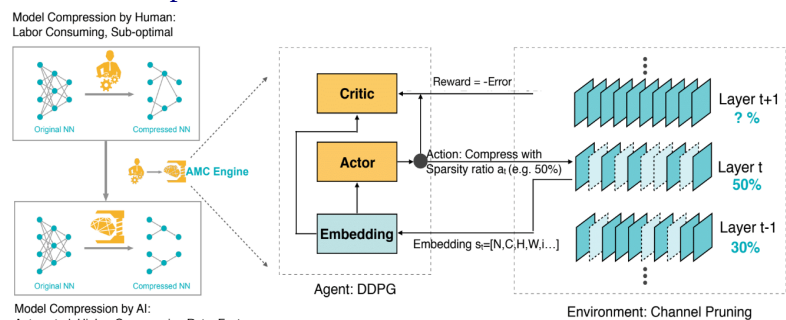
1. Pick layer L_i in the model
 - Prune on the layer L_i with pruning ratio $r \in \{0, 0.1, 0.2, \dots, 0.9\}$
 - Observe the accuracy degrade ΔAcc_i^r for each pruning ratio
2. Repeat process for all layers
3. Pick degradation threshold T such that the overall pruning rate is desired

OPTIMAL? Maybe not! We don't consider interaction between layers.

Automatic Pruning [AMC: AutoML for Model Compression](#)

AMC uses following setup for the reinforcement learning problem:

- State: features including layer indices, channel numbers, kernel sizes, FLOPs, ...
- Action: A continuous number (pruning ratio) $a \in [0, 1]$
- Agent: DDPG agent, since it support continuous action output
- Reward $R = \begin{cases} -Error, & \text{if satisfies constraints} \\ -\infty, & \text{if not} \end{cases}$



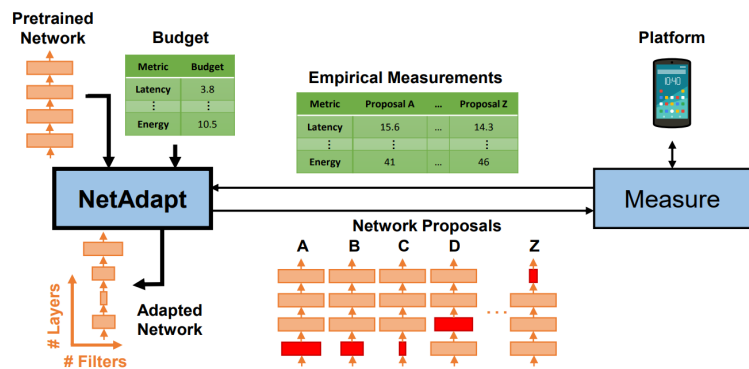
NetAdapt

Find per-layer pruning ratio to meet global resource constraint (e.g. latency energy)

For each iteration, reduce latency by certain amount ΔR

- For each layer L_k (k in A-Z)
 - Prune layer s.t. latency reduction meets ΔR
 - Short-term fine-tune model; measure accuracy after fine tuning
- Choose and prune the layer with highest accuracy

Repeat until total latency reduction satisfies the constraint. Long term fine-tune to recover accuracy.



Finetuning Pruned Neural Networks

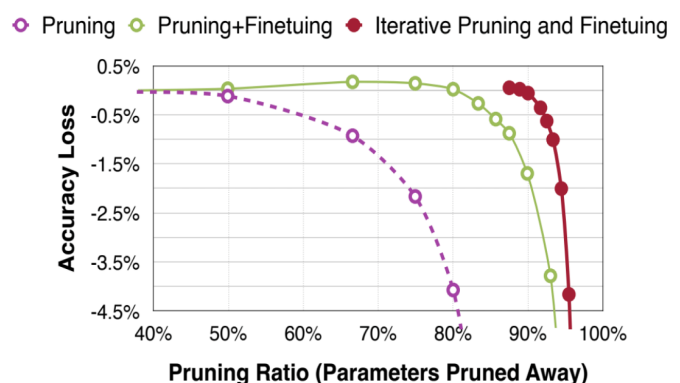
Learning rate for fine-tuning is usually 1/100 or 1/10 of the original learning rate.

Iterative pruning gradually increases the target sparsity in each iteration e.g. boost pruning ratio from 5x to 9x on AlexNet compared to single-step aggressive pruning.

When training neural network or fine-tuning quantized neural network, regularization is added to loss term to (a) penalize non-zero parameters (b) encourage smaller parameters.

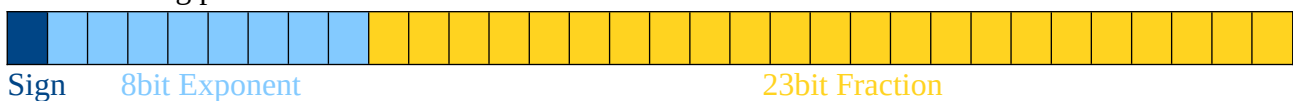
Examples:

- Magnitude-based Fine-grained Pruning applies L2 regularization on weights
- Network Slimming applies smooth L1 regularization on channels scaling factors.



Quantization

32-bit floating point number in IEEE 754



$$(-1)^{sign} \times (1 + \text{Fraction}) \times 2^{\text{Exponent} - 127}$$

$$0.265625 = 1.0625 \times 2^{-2} = (1 + 0.0625) \times 2^{125 - 127}$$



- Subnormal number, when exponent bits are frozen to 0 $(-1)^{sign} \times \text{Fraction} \times 2^{1-127}$
- Normal number, when exponent bits are frozen to 1
 - If signed bit is 0 and fraction bits are all 0, it represent $+\infty$
 - If signed bit is 1 and fraction bits are all 0, it represent $-\infty$
 - If fractions bits are non-zero, it represent NaN (Not a Number)

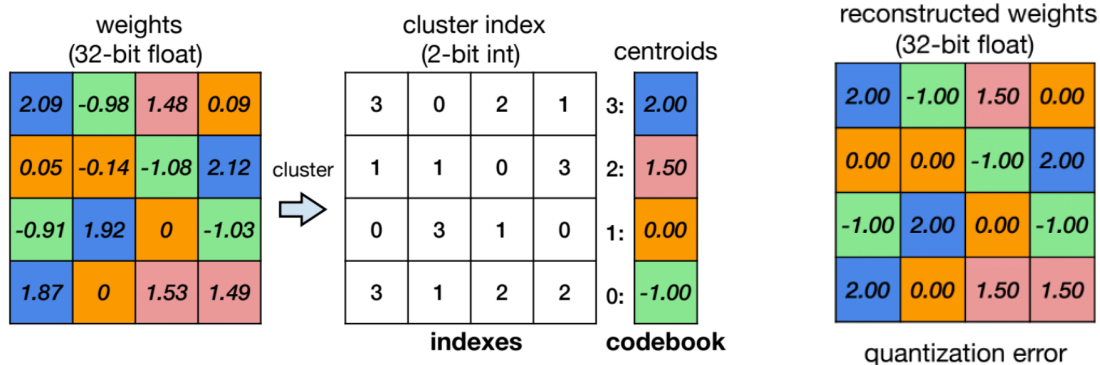
Mantissa is used for precision while exponent is used for dynamic range. Google Brain BF16 representation, Nvidia FP8(E4M3), Nvidia FP8 (E5M2) also comes to picture.

Format	Exponent (bits)	Fraction (bits)
IEEE754 32-bit (IEEE FP32)	8	23
IEEE754 16-bit (IEEE FP16)	5	10
Google Brain Float (BF16)	8	7
Nvidia FP8 (E4M3)	4	3
Nvidia FP8 (E5M2)	5	2

FP8(E4M3) does not have ∞ ,
S.1111.111₂ is used for NaN
(Largest value 448)

FP8(E5M2) does have ∞
(S.11111.00₂) and Nan
(S.11111.XX₂) (Largest value
57334.)

K-Means-based weight Quantization



Storage (4x4 weight matrix):

Initially 32bit x 16 = 512 bit

After quantization: Indexes 2 bit x 16 + codebook 32bits x 4 =
= 160bit

Reduction 3.2x

Assume N bit quantization, and #param M $\gg 2^N$

Initially 32bit x M = 32M bit

After quantization: Indexes N bit x M + codebook 32bits x 4 =
 $\approx NM$ bit

Reduction 32/N

Fine-tuning Quantized weights

First do a group-by operation on gradients.

Reduce them to single value (summing or take mean) multiply them with a learning rate and shifts the centroids.

Summary:

The weights are decompressed using lookup table (i.e. codebook) during runtime inference.

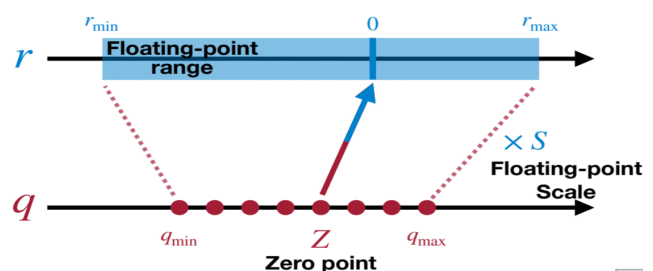
K-means-based weight Quantization only saves storage cost of neural network model. All the computation and memory access are still floating.

Linear Quantization

An affine mapping of integers to real numbers.

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

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



Consider following matrix multiplication

$$Y = WX$$

$$S_Y(q_Y - Z_Y) = S_W(q_W - Z_W) \cdot S_X(q_X - Z_X)$$

$$q_Y = \frac{S_W S_X}{S_Y} (q_W - Z_W)(q_X - Z_X) + Z_Y$$

$$q_Y = \frac{S_W S_X}{S_Y} (q_W q_X - Z_W q_X - Z_X q_W + Z_W Z_X) + Z_Y$$

Use N bit Integer Multiplication and 32-bit Integer Addition/Subtraction.

Pre-compute $-Z_X q_W + Z_W Z_X$ No run-time overhead.

Empirically, the scale $\frac{S_W S_X}{S_Y}$ is always (0,1) $\frac{S_W S_X}{S_Y} = 2^{-n} M_0$, where $M_0 \in [0.5, 1)$. 2^{-n} is simple bit shift operation and multiplication with M_0 is fixed-point multiplication.

$$q_Y = \underbrace{\frac{S_W S_X}{S_Y}}_{\text{Rescale to N-bit int}} (q_W q_X - \underbrace{Z_W q_X - Z_X q_W + Z_W Z_X}_{\text{Precompute}}) + \underbrace{Z_Y}_{\text{N-bit int addition}}$$

If we set zero point $Z=0, Z_W=0$

$$q_Y = \frac{S_W S_X}{S_Y} (q_W q_X - Z_X q_W) + Z_Y$$

Now consider fully connected layer with bias

$$Y = WX + b$$

$$S_Y(q_Y - Z_Y) = S_W(q_W - Z_W) \cdot S_X(q_X - Z_X) + S_b(q_b - Z_b)$$

$\downarrow Z_W = 0$

$$S_Y(q_Y - Z_Y) = S_W S_X (q_W q_X - Z_X q_W) + S_b(q_b - Z_b)$$

$\downarrow Z_b = 0, S_b = S_W S_X$

$$S_Y(q_Y - Z_Y) = S_W S_X (q_W q_X - Z_X q_W + q_b)$$

$$q_Y = \frac{S_W S_X}{S_Y} (q_W q_X + \underbrace{q_b - Z_X q_W}_{\text{Precompute}}) + Z_Y$$

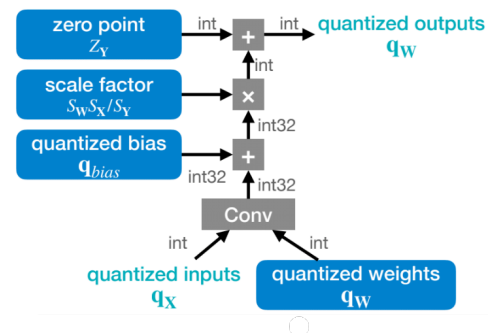
$\downarrow q_{bias} = q_b - Z_X q_W$

$$q_Y = \frac{S_W S_X}{S_Y} (q_W q_X + q_{bias}) + Z_Y$$

Note: q_b and q_{bias} are 32 bits

For Convolution, results will be the same

$$q_Y = \frac{S_W S_X}{S_Y} (\text{Conv}(q_W q_X + q_{bias})) + Z_Y, \text{ both } q_b \text{ and } q_{bias} \text{ are 32 bits}$$



		K-Means-based Quantization	Linear Quantization
Storage	Floating-Point Weights	Integer Weights, Floating Point Codebook	Integer Arithmetic
Computation	Floating-Point Arithmetic	Floating Point Arithmetic	Integer Arithmetic

