Appendix A. Quantization Error Measurement

The difference of mean value before and after quantization is $\frac{m_d}{m_{\hat{d}}} = \frac{\int_a^b P(x)xdx}{a\int_a^b P(x)dx + b\int_a^b P(x)dx}$. We use $P(x) = kx + o$ to approximate the local value between $[a, b]$ with $b < -\frac{o}{k}, k < 0$, and assign $C = \frac{1}{4}k(a+b)^2 + \frac{o(a+b)}{2}$, then we have:

$$\begin{aligned}\int_a^b P(x)xdx &= \int_a^b (kx^2 + ox)dx \\ &= \frac{1}{3}kx^3 + \frac{1}{2}kx^2 \Big|_a^b \\ &= (\frac{1}{3}k(a^2 + b^2 + ab) + \frac{o}{2}(a+b))(b-a)\end{aligned} \tag{1}$$

$$\begin{aligned}a\int_a^c P(x)dx + b\int_c^b P(x)dx &= a\int_a^c (kx+o)dx + b\int_c^b (kx+o)dx \\ &= a\,(\frac{1}{2}kx^2 + ox)\Big|_a^c + b\,(\frac{1}{2}kx^2 + ox)\Big|_c^b \\ &= (\frac{1}{8}k(3a^2 + 3b^2 + 2ab) + \frac{o}{2}(a+b))(b-a)\end{aligned} \tag{2}$$

$$\frac{m_d}{m_{\hat{d}}} = \frac{\frac{1}{3}k(a^2 + b^2 + ab) + \frac{o}{2}(a+b)}{\frac{1}{8}k(3a^2 + 3b^2 + 2ab) + \frac{o}{2}(a+b)} \tag{3}$$

$$\frac{1}{3}k(a^2 + b^2 + ab) - \frac{1}{8}k(3a^2 + 3b^2 + 2ab) = -\frac{1}{24}(a+b)^2k > 0 \tag{4}$$

Considering Equation. 4 in Equation. 3 , we have $\frac{m_d}{m_{\hat{d}}} > 1$. Denote $A = a + b$ and $B = b - a$, then Equation. 3 becomes:

$$\frac{m_d}{m_{\hat{d}}} = \frac{\frac{1}{4}kA^2 + \frac{1}{2}oA + \frac{1}{12}kB^2}{\frac{1}{4}kA^2 + \frac{1}{2}oA + \frac{1}{8}kB^2} \tag{5}$$

As $b < -\frac{o}{k}$, so $\frac{k}{o} > -\frac{1}{b}$

$$\begin{aligned}C &= \frac{1}{4}kA^2 + \frac{1}{2}oA \\ C = \frac{1}{4}Ao(\frac{k}{o}A + 2) &> \frac{1}{4}A\frac{b-a}{b} > 0\end{aligned} \tag{6}$$

Therefore,

$$\begin{aligned}\frac{m_d}{m_{\hat{d}}} &= \frac{C + \frac{1}{12}kB^2}{C + \frac{1}{8}kB^2} \\ &= 1 - \frac{1/24}{\frac{C}{B^2k} + 1/8} \\ &= 1 + \frac{1/24}{\frac{C}{(b-a)^2(-k)} - 1/8}\end{aligned} \tag{7}$$

Appendix B. Quantification Method

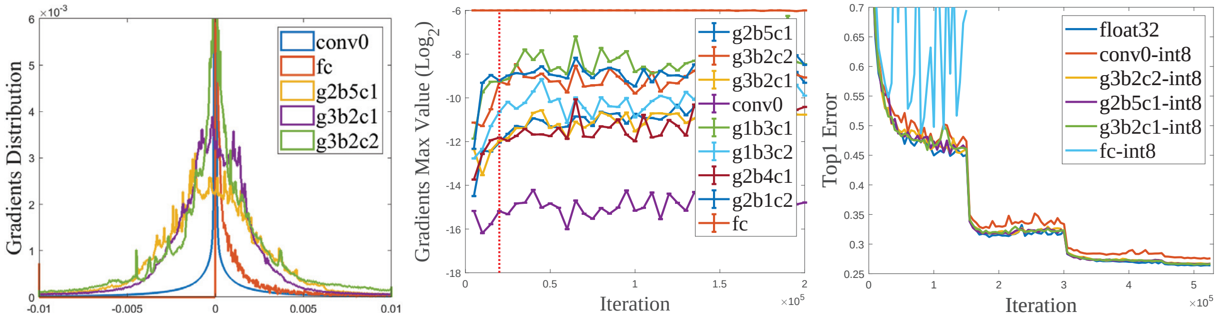
A fixed-point number consists of a sign bit, $(n-1)$-bit integer, and a global quantization resolution r relating to fixed-point position s. Before quantization, the maximum absolute data is Z. The representation data range, bit-width and quantization resolution are inter-dependent, as $Range \approx r \times 2^n$. The quantization resolution is calculated as in Table. 1 column 2. Suppose F_x is the floating point representation of x and I_x is the fixed-point representation of x, and $\hat{F_x}$ is the approximation of F_x, as $\hat{F_{x_1}} = I_{x_1} \times r_1$, $\hat{F_{x_2}} = I_{x_2} \times r_2$, the multiplication between numbers becomes:

$$F_{x_1} \times F_{x_2} \approx \hat{F_{x_1}} \times \hat{F_{x_2}} = r_1 \times r_2 \times I_{x_1} \times I_{x_2} \tag{8}$$

Table 1: Quantization method

Quantization Function	Quantization Resolution	Fixed-Point Data Range
$I_x = round(\frac{F_x}{r})$	$r = 2^s = 2^{ceil(log_2(\frac{Z}{2^{n-1}-1}))}$	$[-r(2^{n-1}), r(2^{n-1}-1)]$

Appendix C. Observations on Other Network



(a) Activation gradient distribution.

(b) Activation gradient evolution.

(c) Training curve of AlexNet.

Figure 1: Observations on ResNet34.

As shown in Figure. 1, for ResNet34 int8 is enough to quantify the activation gradient of g3b2c2, g2b5c1 and g3b2c1, however, int8 for fc and conv0 either not converges or introduces accuracy drop, conv0 and fc have large variance. These observations are consistent with the observation on AlexNet. In conclusion, data with large variance requires large bit-width, thus the quantization parameters should be dynamically determined by the data distribution.

Appendix D. Extra Evaluation of Error Measurement

Figure. 2 shows the linear correlation between ResNet50 accuracy and several error metrics. Our proposed quantization error measurement M1 has the highest correlation score (0.85 for ResNet50) with the network-level accuracy.

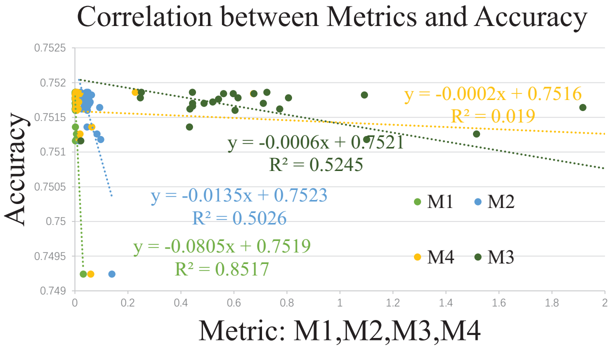


Figure 2: Correlation between ResNet50 accuracy a and quantization error measurement M.

Appendix E. Adaptive Bit-width for Weight and Activation

Here we show the results of adaptive lower bit-width (i.e., int4) for forward-pass. All the data of AlexNet and ResNet18 are quantified into int4 at start. Then, the bit-width of different layer is automatically increased by the proposed QEM and QPA. As shown in Table. 2, 62.5% of linear layers of AlexNet and 36.3% of linear layers of ResNet18 involve int4 multiplication [1].

[1] Weight of conv0, fc0, fc2 on AlexNet are int4, activation of conv0, conv1, conv2 on AlexNet are int4. Weight of conv1, res2a_1, res2b_2a, res2b_2b on ResNet18 are int4, activation of conv1, res2b_2a, res2b_2b, res3a_1, res3a_2a, res3a_2b, fc on ResNet18 are int4.

The final mixed precision models have small accuracy losses (0.3%˜0.9%). Quantify the back propagation is our primary pursuit, but the proposed QEM and QPA can also be extended to low-bit inference (e.g., binary or ternary), which will be our future work.

Table 2: 4-bit fine-tuning

Classification	**Baseline(float32)**	**Adaptive**	**Weight Bit-width**		**Activation Bit-width**	
Network	**top1 accuracy**	**top1 accuracy**	int4	int8	int4	int8
AlexNet	58.0	57.7	37.5%	62.5%	50%	50%
ResNet18	67.3	66.4	19%	81%	33.3%	66.7%

Appendix F. Training loss convergence

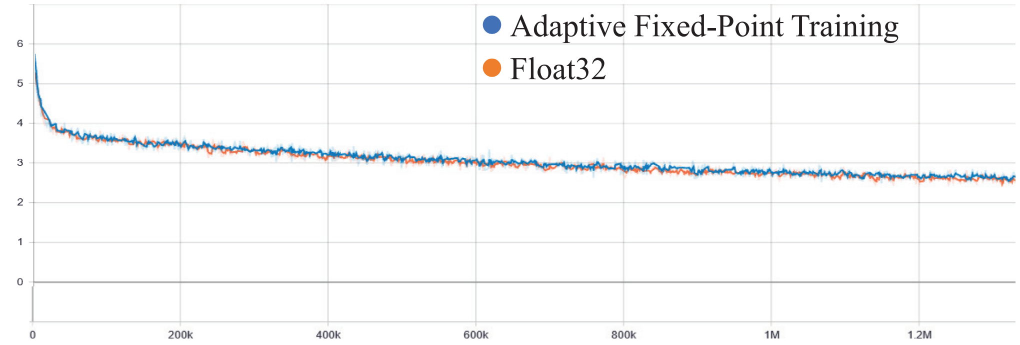


Figure 3: Training loss curve for MobileNet v2.

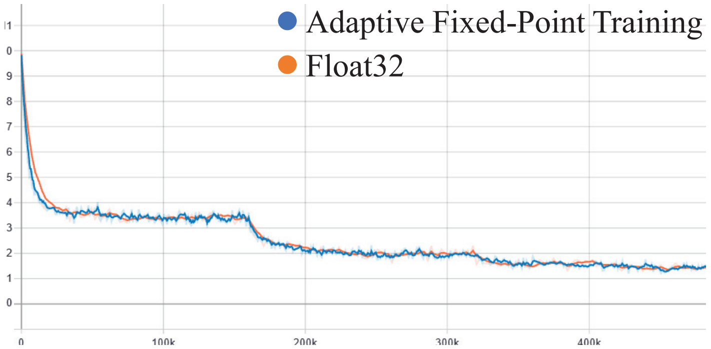


Figure 4: Training loss curve for ResNet50.

The training loss curves of for Mobilenet v2 and ResNet50 are shown in Figure. 3 and Figure. 4. Adaptive Fixed-Point Training has the same convergence speed as float32 training.

Appendix G. Speedup over int16

There is 1.3 times speedup over int16 on CPU for AlexNet (1.13 times speedup for backward and 1.7 times speedup for forward). The int16 x int8 in our method is implemented as int16 x int16 on Xeon Gold 6154. With flexible arithmetic operations like int16 x int8 on future hardware, higher training speedup is promising.