RDO-Q: Extremely Fine-Grained Channel-Wise Quantization via Rate-Distortion Optimization —Supplementary Material

Zhe Wang[1], Jie Lin[1], Xue Geng[1], Mohamed M. Sabry Aly[2], and Vijay Chandrasekhar[1,2]

[1] Institute for Infocomm Research, A*STAR, Singapore 138632
wangz@i2r.a-star.edu.sg, jie.dellinger@gmail.com, geng_xue@i2r.a-star.edu.sg, vijay.cmu@gmail.com
[2] Nanyang Technological University, 50 Nanyang Ave, Singapore 639798
msabry@ntu.edu.sg

Abstract. In this supplementary materials, we provide the analysis for the additivity property of output distortion. We show a mathematical derivation for the additivity property by linearizing the output distortion using Taylor series expansion.

1 Additivity of Output Distortion

The output distortion δ, caused by quantizing all weight channels and activation layers, equals the sum of all output distortion due to the quantization of each individual weight channel and activation layer

$$\delta = \sum_{i=1}^{l} \sum_{j=1}^{n_i} \delta_{i,j}^{w} + \sum_{i=1}^{l} \delta_i^a \tag{1}$$

if the neural network is continuously differentiable in every layer, the quantization errors can be considered as small deviations distributed with zero mean.

Proof. We first define the main notations. Let

$$\mathcal{F}(W_{1,1}, ..., W_{1,n_1}, ..., W_{l,1}, ..., W_{l,n_l})$$

denote a neural network and

$$\widetilde{\mathcal{F}}(W_{1,1}, ..., W_{1,n_1}, ..., W_{l,1}, ..., W_{l,n_l}, s_1, ..., s_l)$$

denote a modified neural network of $\mathcal{F}$ where an element-wise add layer with parameter s_i is followed for each activation a_i (see Fig. 1). Based on this definition, we have

$$\mathcal{F}(W_{1,1}, ..., W_{l,n_l}) = \widetilde{\mathcal{F}}(W_{1,1}, ..., W_{l,n_l}, 0, ..., 0) \tag{2}$$

Define two variables X_0 and ΔX, where

$$X_0 = \big(W_{1,1}, ..., W_{l,n_l}, 0, ..., 0\big)$$

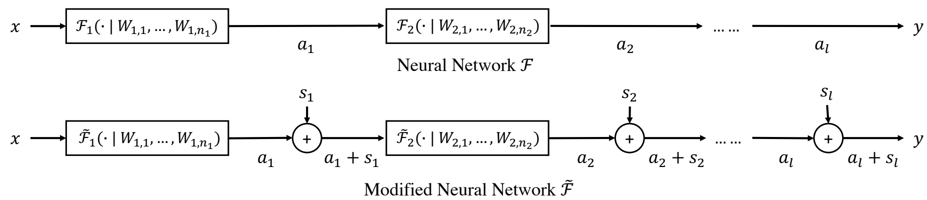

Fig. 1. Examples of a neural network $\mathcal{F}$ and a modified neural network $\widetilde{\mathcal{F}}$.

$$\varDelta X = \left(\varDelta W_{1,1}, ..., \varDelta W_{l,n_l}, \varDelta s_1, ..., \varDelta s_l\right)$$

Assume that the quantization error can be considered as small deviation. We apply the Taylor series expansion up to first order term on $\widetilde{\mathcal{F}}$ at X_0,

$$\begin{aligned} \widetilde{\mathcal{F}}(X_0 + \varDelta X) - \widetilde{\mathcal{F}}(X_0) =& \sum_{i,j} \frac{\partial \widetilde{\mathcal{F}}}{\partial W_{i,j}} \cdot \varDelta W_{i,j} \\ &+ \sum_{i} \frac{\partial \widetilde{\mathcal{F}}}{\partial s_i} \cdot \varDelta s_i. \end{aligned} \tag{3}$$

Then $\|\widetilde{\mathcal{F}}(X_0 + \varDelta X) - \widetilde{\mathcal{F}}(X_0)\|^2$ can be written as

$$\begin{aligned} &\Big(\sum_{i,j} \varDelta W_{i,j}{}^{\top} \cdot \frac{\partial \widetilde{\mathcal{F}}}{\partial W_{i,j}}^{\top} + \sum_{i} \varDelta s_i{}^{\top} \cdot \frac{\partial \widetilde{\mathcal{F}}}{\partial s_i}^{\top}\Big) \\ &\cdot\Big(\sum_{i,j} \frac{\partial \widetilde{\mathcal{F}}}{\partial W_{i,j}} \cdot \varDelta W_{i,j} + \sum_{i} \frac{\partial \widetilde{\mathcal{F}}}{\partial s_i} \cdot \varDelta s_i\Big) \end{aligned} \tag{4}$$

Because quantization errors in different layers are independently distributed with zero mean, the cross terms of (4) disappear when taking the expectation. That is:

$$\begin{aligned} &E(\varDelta W_{i,j}{}^{\top} \cdot \frac{\partial \widetilde{\mathcal{F}}}{\partial W_{i,j}}^{\top} \cdot \frac{\partial \widetilde{\mathcal{F}}}{\partial W_{i,j}} \cdot \varDelta W_{i,j}) \\ =&E(\varDelta W_{i,j}{}^{\top}) \cdot \frac{\partial \widetilde{\mathcal{F}}}{\partial W_{i,j}}^{\top} \cdot \frac{\partial \widetilde{\mathcal{F}}}{\partial W_{i,j}} \cdot E(\varDelta W_{i,j}) = 0 \end{aligned} \tag{5}$$

as is the case also for the cross products between $W_{i,j}$ and s_i (all i, j), and s_i and s_j ($i \neq j$). Then, we can obtain

$$\begin{aligned} &E(\|\widetilde{\mathcal{F}}(X_0 + \varDelta X) - \widetilde{\mathcal{F}}(X_0)\|^2) \\ =&\sum_{i,j} E\Big(\|\frac{\partial \widetilde{\mathcal{F}}}{\partial W_{i,j}} \cdot \varDelta W_{i,j}\|^2\Big) + \sum_{i} E\Big(\|\frac{\partial \widetilde{\mathcal{F}}}{\partial s_i} \cdot \varDelta s_i\|^2\Big) \end{aligned} \tag{6}$$

Eq. (6) is the result we want because, again, according to the Taylor series expansion up to first order term,

$$\begin{aligned}\frac{\partial \widetilde{\mathcal{F}}}{\partial W_{i,j}} \cdot \Delta W_{i,j} = &\widetilde{\mathcal{F}}(..., W_{i,j} + \Delta W_{i,j}, ..., W_{l,n_l}, 0, ...) \\ &- \widetilde{\mathcal{F}}(..., W_{i,j}, ..., W_{l,n_l}, 0, ...)\end{aligned} \tag{7}$$

$$\begin{aligned}\frac{\partial \widetilde{\mathcal{F}}}{\partial s_i} \cdot \Delta s_i = &\widetilde{\mathcal{F}}(W_{1,1}, ..., W_{l,n_l}, 0, ..., \Delta s_i, ...) \\ &- \widetilde{\mathcal{F}}(W_{1,1}, ..., W_{l,n_l}, 0, ..., 0, ...)\end{aligned} \tag{8}$$

After dividing both sides of (6) by the dimensionality of the output vector of the neural network, the left side becomes δ and the right side becomes the sum of all output distortion due to the quantization of each individual weight channel and activation layer.