# Post Training Quantization. Flexible continuous modification for SOTA post training quantization methods to make them lossless.

A Preprint

David S. Hippocampus*
Department of Computer Science
Cranberry-Lemon University
Pittsburgh, PA 15213
hippo@cs.cranberry-lemon.edu

Elias D. Striatum
Department of Electrical Engineering
Mount-Sheikh University
Santa Narimana, Levand
stariate@ee.mount-sheikh.edu

## Abstract

Neural network quantization gives the opportunity to inference large models on resource constrained devices. Post training quantization methods have became popular, as they are simple to use. They don't require whole model retraining and use only small calibration set to calculate quantization parameters. However, these methods show significant accuracy decrease on low-bit quantization. There are methods that allow to increase the accuracy of model by increasing its computational complexity. In this paper, we propose a continuous modification for these methods, and find a reasonable compromise between computational complexity and quantization efficiency.

Keywords First keyword · Second keyword · More

## 1 Introduction

Deep Neural Networks(DNN) are applicable to wide range of tasks nowadays. Despite showing the great performance on these tasks, state-of-the-art models require high computational resources. There is a need to run large models on power-limited devices such as smartphones. Many different methods were proposed for model compression. In this paper, we concentrate on quantization method.

The quantization is a process of mapping real numbers to the low-precision discrete values. There are two main types of quantization methods: quantization aware training(QAT) and post training quantization(PTQ). Quantization aware training shows great results, however, it requires the whole model retraining. Hence, this method is not applicable in some real-life cases, as if training data is not available or computational resources are limited. Unlike QAT, post training quantization usually uses only an unlabeled calibration set for setting up quantization. Current post quantization methods are not such efficient as quantization aware training. However, post training quantization is a promising technique and therefore should be explored further.

The goal of post training quantization is to find optimal quantization parameters having only small set of data. The main problem of this technique is that quantization errors of layers can be amplified by deeper layers. Quantization errors can accumulate layer by layer and lead to accuracy degradation. Quantization accuracy degradation is explored in the [Yury Nahshan, 2020] article, which explaines why low-bit post-training quantization is a quite challenging task.

Most of post training quantization methods quantize model parameters and data by minimizing the error between quantized and the original model layers outputs. The recent post quantization techniques [Itay Hubara, 2021, Yuhang Li, 2021] made a progress towards low-bit post training quantization, considering previous

*Use footnote for providing further information about author (webpage, alternative address)—not for acknowledging funding agencies.

layers errors during quantization. However, these methods leave model structure without changes and don't consider improving accuracy of quantized model by complicating its structure.

In this work, we study ways to improve quantized model accuracy by making model more complex. Paper [Xinghcao Liu, 2021] uses the idea of approximating model weights as a sum of low-precision values. Our paper suggests a modification to this method. There are two main goals of this work. Firstly, we would like to propose a method to make post training quantization lossless. This is relevant to situations when computational device support only low bit data types. Second approach of this paper is to find a trade-off between model complexity and quantization bits, allowing to compress model for resource constrained devices.

## 2 Problem statement section

In this article, we use uniform quantization. Given value to quantize $v$ , the maximum and minimum quantization value $Q_{max}$ and $Q_{min}$ and quantization step size $\Delta$, quantizer computes integer representation of a data $\bar{v}$:

$$\bar{v} = \begin{cases} -Q_{min}, \text{if} \frac{v}{\Delta} \leq -Q_{min} \\ \lfloor \frac{v}{\Delta} \rceil, \text{if} \frac{v}{\Delta} \in [-Q_{min}, Q_{max}] \\ Q_{max}, \text{if} \frac{v}{\Delta} \geq Q_{max} \end{cases} .$$

To get representation of the same scale, $\bar{v}$ is multiplied by $\Delta$:

$$\hat{v} = \bar{v} * \Delta.$$

Let's suppose that that we calculate the quantization of each parameter $W$ $K_W$ times, denote $K = (K_{W_1}, ...K_{W_N})$, where $W_1, ..., W_N$ - model parameters. Also let $\Delta$ be a vector containing quantization parameters for all weights and data.

The goal of our work is to quantize model M without significant performance degradation. We will achieve this by making outputs of $Q(M, \Delta, K)$ similar to the outputs of M, where $Q(M, \Delta, K)$ is a denotion for quantized model M with parameters $K$, $\Delta$. For comparing models outputs, we will use some loss function $L(M1, M2, dataset)$.

Let's denote model M complexity as $P(M)$.

Then, we want to minimize $L(M, Q(M, \Delta, K), dataset)$ for given model M, calibration dataset $dataset$ and some complexity limit $P_0$:

$argmin_{K,\Delta}\{L(M, Q(M, \Delta, K), dataset), P(M, K, \Delta) \leq P_0\}$

### 2.1 Headings: second level

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

$$\xi_{ij}(t) = P(x_t = i, x_{t+1} = j|y, v, w; \theta) = \frac{\alpha_i(t) a_{ij}^{w_t} \beta_j(t+1) b_j^{v_{t+1}}(y_{t+1})}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i(t) a_{ij}^{w_t} \beta_j(t+1) b_j^{v_{t+1}}(y_{t+1})} \tag{1}$$

#### 2.1.1 Headings: third level

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.


Рис. 1: Sample figure caption.

Paragraph Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

## 3 Examples of citations, figures, tables, references

### 3.1 Citations

Citations use `natbib`. The documentation may be found at

`http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf`

Here is an example usage of the two main commands (`citet` and `citep`): Some people thought a thing [??] but other people thought something else [?]. Many people have speculated that if we knew exactly why ? thought this...

### 3.2 Figures

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi. See Figure 1. Here is how you add footnotes. [2] Sed feugiat. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Ut pellentesque augue sed urna. Vestibulum diam eros, fringilla et, consectetuer eu, nonummy id, sapien. Nullam at lectus. In sagittis ultrices mauris. Curabitur malesuada erat sit amet massa. Fusce blandit. Aliquam erat volutpat. Aliquam euismod. Aenean vel lectus. Nunc imperdiet justo nec dolor.

[2]Sample of the first footnote.

Таблица 1: Sample table title

| Part | | |
|---|---|---|
| Name | Description | Size ($\mu$m) |
| Dendrite | Input terminal | $\sim$100 |
| Axon | Output terminal | $\sim$10 |
| Soma | Cell body | up to $10^6$ |

### 3.3 Tables

See awesome Table 1.

The documentation for `booktabs` ('Publication quality tables in LaTeX') is available from:

`https://www.ctan.org/pkg/booktabs`

### 3.4 Lists

- Lorem ipsum dolor sit amet
- consectetur adipiscing elit.
- Aliquam dignissim blandit est, in dictum tortor gravida eget. In ac rutrum magna.

## Список литературы

Chaim Baskin Evgenii Zheltonozhskii Ron Banner Alex M. Bronstein Avi Mendelson Yury Nahshan, Brian Chmiel. Loss aware post-training quantization. 2020.

Yair Hanani Ron Banner Daniel Soudry Itay Hubara, Yury Nahshan. Accurate post training quantization with small calibration sets. In Proceedings of the 38th International Conference on Machine Learning. PMLR, 2021.

Xu Tan Yang Yang Peng Hu Qi Zhang Fengwei Yu Wei Wang Shi Gu Yuhang Li, Ruihao Gong. Brecq: Pushing the limit of post-training quantization by block reconstruction. University of Electronic Science and Technology of China, SenseTime Research, 2021.

Dengyong Zhou Qiang Liu Xinghcao Liu, Mao Ye1. Post-training quantization with multiple points: Mixed precision without mixed precision. 2021.