Post Training Quantization. Flexible continuous modification for SOTA post training quantization methods to make them lossless. 1

A Preprint

Zharikov Ilya
Moscow Institute of Physics and Technology

Sedova Anna
Moscow Institute of Physics and Technology

Abstract

Neural network quantization gives the opportunity to inference large models on resource constrained devices. Post-Training Quantization(PTQ) methods have became popular, as they are simple and fast to use. They do not require whole model retraining and use only small calibration set to calculate quantization parameters. However, these methods show significant accuracy decrease on low-bit setting. 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 trade-off between computational complexity and performance.

Keywords Deep Learning · Model Compression · Post-Training Quantization

1 Introduction

ToDo: Need to be extended in terms related works further.

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. 1 2 3

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). 4 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-training quantization methods are not such efficient as quantization aware training. However, post-training quantization is a promising technique and therefore should be explored further. 5

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 paper Yury Nahshan [2020] article, which explains 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-training quantization techniques [Itay Hubara, 2021, Yuhang Li, 2021] made a progress towards low-bit post-training quantization, considering previous 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.

1. Очень длинное название. Понятно, что вариант рабочий, но имеет смысл подумать над более емким названием. И порядок авторов другой

2. high compuitational resources - нужна ссылка на литературу, подкрепляющую утверждение

3. аналогично, см. пункт 2

3. я бы добавил и сюда какую-нибудь ссылку на общую литературу по квантизации

4. сюда тоже нужны ссылки на литературу

5. и сюда

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. 6

2 Problem statement

ToDo: Need to be slightly reformulated during theory week.

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 α, quantizer computes integer representation of a data $\bar{v}$:

$$\bar{v} = \begin{cases} -Q_{min}, \text{if} \frac{v}{\alpha} \leq -Q_{min} \\ \lfloor \frac{v}{\alpha} \rceil, \text{if} \frac{v}{\alpha} \in [-Q_{min}, Q_{max}] \\ Q_{max}, \text{if} \frac{v}{\alpha} \geq Q_{max} \end{cases} .$$ 7

To get representation of the same scale, $\bar{v}$ is multiplied by α: 8

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

~~Let~~'s suppose that model has n parameters $W_1, ..., W_n$, then let~~'s~~ denote the model consisting of these parameters as $M(W_1, ..., W_n)$. Also let quantized model parameters be denoted as $Q(W_1, \alpha_1), ..., Q(W_n, \alpha_n)$.

The goal of our work is to quantize model M without significant performance degradation. We will achieve this by making outputs of $M(Q(W_1, \alpha_1), ..., Q(W_n, \alpha_n))$ similar to the outputs of $M(W_1, ..., W_n)$. Let's denote model M complexity as $P(M)$, model quality as $F(M)$. 9

Then, we want to maximize $F(M(Q(W_1, \alpha_1), ..., Q(W_n, \alpha_n)))$ for given model $M(W_1, ..., W_n)$ and some complexity limit P_0:

$$\arg \max_{\alpha_1, ..., \alpha_n} \{F(M(Q(W_1, \alpha_1), ..., Q(W_n, \alpha_n)))\ , P(M(Q(W_1, \alpha_1), ..., Q(W_n, \alpha_n))) \leq P_0\}$$

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.

6. Нужен текст о том, как вы проводите эксперимент (кратко, 2-3 предложения)

7. Кажется, потерялись пробелы между if и формулами

8. Это произведение двух числел? почему бы не сделать \alpha\bar{v}?

9. Нужно подробнее расписать что вы имеете ввиду в работе под quality и complexity