REPRODUCIBILITY CHALLENGE : AND THE BIT GOES DOWN: REVISITING THE QUANTIZATION OF NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Quantization is one of the popular approaches to compress the deep neural networks. Various quantization methods had been proposed, and *And the Bit Goes Down: Revisiting the Quantization of Neural Networks* Stock et al. (2020) proposed a vector quantization method that preserves the quality of the reconstruction of the network outputs rather than their weight. This paper suggested a codebook idea that stores the compressed weights. However, since this paper run the experiments only on the ImageNet dataset. Furthermore, not only this paper but also most quantization method papers focus on the ImageNet dataset. We were curious about whether this method works on other datasets. In this paper, we run the proposed method on Stanford Dogs and Oxford 102 Flowers dataset, which contains similar images which have small interclass variances. Comprehensive experimental results demonstrate that the original method does not work on Stanford Dogs and Oxford 102 Flowers dataset.

1 INTRODUCTION

Our chosen paper for the reproducibility challenge is *And the Bit Goes Down: Revisiting the Quantization of Neural Networks* Stock et al. (2020). This paper introduced the vector quantization method that aims at preserving the quality of the reconstruction of the network outputs rather than their weight. It also deals with the problem of reducing the memory of convolutional network architectures. Also, they leverage the spatial redundancy of the unique information in standard convolutional filters Denton et al. (2014). This paper only requires unlabelled data at quantization time and uses codebooks to store the compressed weights which are useful for inference on CPU. Experiments in this paper demonstrated high performance for the ResNet-50 model, with a memory size of 5 MB, which means that the compression factor is 20x. For Mask R-CNN, the memory is about 6.65 MB, which has a compression factor of 26x. In object detection using the ResNet-50 model, the original top-1 accuracy is 76.15%, which shows the corresponding accuracy even after compression. The original code is available at `https://github.com/facebookresearch/kill-the-bits`.

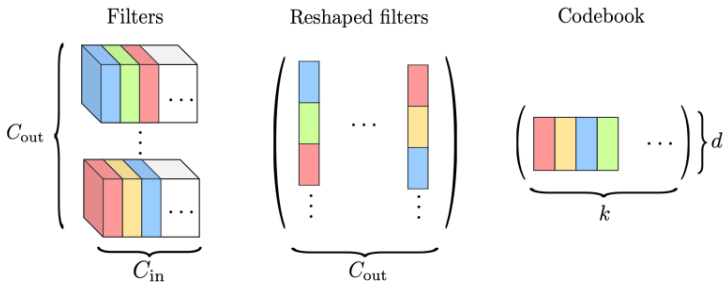

Figure 1: This is the process of quantizing the filter and converting it into a codebook.

We are mainly interested in various model compression and low-power papers. There are several models of compressing tactics such as pruning and quantization. Most of the pruning papers have done several experiments for several datasets, but various quantization papers have done experiments only on ImageNet dataset Deng et al. (2009). Also in this paper, the author had done experiments only on the ImageNet benchmark.

It is known that if a specific algorithm works for large datasets, it will also work for small datasets. But it might not work for datasets that have images of the same class. Therefore, we want to check if the algorithm proposed by this paper is also working for other datasets. Especially, we will use Oxford 102 Flowers Nilsback & Zisserman (2008) dataset and Stanford Dogs dataset Khosla et al. (2011). As seen in Figure 2 and Figure 3, the general difficulty of these datasets is they contain images that have visually small interclass variances. This kind of dataset will consider each of the various features of the image as an important factor, but by applying product quantization and codebook idea, those features might be lost. In addition, this paper reduced the memory footprint by using compression, and quantization was also performed. Therefore, we plan to observe whether the memory footprint is also reduced for the datasets mentioned above.

Figure 2: Random example images from Stanford Dogs dataset.

Figure 3: Random example images from Stanford Dogs dataset.

2 REPRODUCIBILITY

We evaluate the original method with Stanford Dogs dataset and Oxford 102 Flowers dataset. To compare the compressed model with the original one, we get the original result by applying transfer learning to pre-trained ResNet models. With pre-trained ResNet-18 and ResNet-50

networks, we add a simple custom classifier. Custom classifier replaced fully connected layer of original models. Replaced fully connected layer is simple since it only contains nn.Linear and F.log_softmax function. We will just simply call those modified models ResNet-18 and ResNet-50. After training the Standford Dogs and Oxford 102 Flowers dataset, the top-1 accuracy of each of them result in 82.53% and 78.53% for ResNet-18, and 87.91% and 86.86% for ResNet-50.

According to the original paper, for ResNet-18, the results after applying this method are 4% lower than the original top-1 accuracy, and for ResNet-50, 2.5% lower than the original one. Our approach is to find out whether the algorithm of the paper works for other datasets. If the accuracy drops in 4% for ResNet-18 and ResNet-50, the original method works on those two datasets.

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

We apply the original quantization method on ResNet-18 and ResNet-50 pre-trained on the ImageNet dataset and replaced the fully connected layer. We use the Standford Dog dataset and Oxford 102 Flowers dataset. The pre-trained models are taken from PyTorch torchvision. We run the experiments on 2 24 GB RTX 3090 GPU. Experiments on the Standford Dog dataset take 2 days and on Oxford 102 Flowers dataset takes 1 day.

The original paper includes two compression regimes, large block sizes, and small block sizes. But, because of lack of time and resources, our reproducibility challenge only runs small block sizes.

3.2 IMAGE CLASSIFICATION RESULTS

We report the results of the quantization method applied to ResNet-18 and ResNet-50 models on the Standford Dogs and Oxford 102 Flowers dataset. Table 1 shows the results of the Standford Dogs dataset, and Table 2 shows the results of the Oxford 102 Flowers dataset. We include the compression ratio between the original model size and compressed model size and the Top-1 accuracy of the original model and compressed model.

Table 1: Results for ResNet-18 and ResNet-50 models on Stanford Dogs for k=256 centroids.

Model (original Top-1)	Compression	Size ratio	Model size	Top-1 (%)
ResNet-18 (82.53%)	Small blocks	30.91×	1.38 MB	3.77
ResNet-50 (87.91%)	Small blocks	19.85×	4.57 MB	10.91

Table 2: Results for ResNet-18 and ResNet-50 models on Oxford 102 Flowers for k=256 centroids.

Model (original Top-1)	Compression	Size ratio	Model size	Top-1 (%)
ResNet-18 (78.53%)	Small blocks	30.89×	1.39 MB	7.25
ResNet-50 (86.86%)	Small blocks	19.87×	4.55 MB	10

Size ratio in the Tables shows that quantization method reduced the memory footprint. However, the top-1 accuracy for Stanford Dogs and Oxford 102 Flowers is not so good. The accuracy of ResNet-18 is less than 10% and of ResNet-50 is about 10%. These accuracy results show that the quantization method proposed from *And the Bit Goes Down: Revisiting the Quantization of Neural*

Networks Stock et al. (2020) does not work well on datasets that have small differences between datasets.

4 Discussion and Conclusion

The original paper mentioned that their method can be adapted to simultaneously compress and transfer ResNets trained on ImageNet to other domains. This is equal to what we have done for the reproducibility challenge. This reproducibility challenge shows that even models that work well on large datasets for various classes may not work well in the same class. We transfer ResNet models pre-trained on the ImageNet dataset to other datasets. However, unlike what the original paper mentioned, the quantization method did not work. We guess the reason why it did not work is about not adjusting hyperparameters. In addition, since the classes of Oxford flowers 102 dataset eventually, all become flowers and the Stanford dog dataset's classes mean dogs, the original paper is useful for classifying flowers and dogs, but we think the model may have difficulty in classifying the kind of flowers and dogs.

To conclude, we applaud the author for proposing a new quantization idea, which might be improved and applied to various models and datasets.

References

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

Emily Denton, Wojciech Zaremba, Joan Bruna, Yann LeCun, and Rob Fergus. Exploiting linear structure within convolutional networks for efficient evaluation. *CoRR*, abs/1404.0736, 2014. URL `http://arxiv.org/abs/1404.0736`.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics Image Processing (ICVGIP)*, 2008.

Pierre Stock, Armand Joulin, Rémi Gribonval, Benjamin Graham, and Hervé Jégou. And the bit goes down: Revisiting the quantization of neural networks. In *International Conference on Learning Representations (ICLR)*, 2020.