Video Encoder-Based Compression for Convolutional Neural Network

Seunghwan Kim
Sungkyunkwan University, Korea
Email: seunghwankim@skku.edu

Eun-Soo Park
Sungkyunkwan University, Korea
Email: espark804@skku.edu

Eun-Seok Ryu
Sungkyunkwan University, Korea
Email: esryu@skku.edu

***Abstract*—Lightweight Convolutional Neural Network (CNN) for mobile devices has been researched in many ways. Much research have achieved great space-saving and reduced computational complexity, maintaining the accuracy. These methods have re-training or fine-tuning step to recover accuracy. In real situation, the re-training steps have some difficulties with computing resource or data access. On these scores, we propose the CNN compression method without fine-tuning step for mobile devices. As most mobile devices includes hardware for video coding, proposing method can be used in mobile devices. In this study, we evaluate our method with different datasets and compare with other methods.**

I. Introduction

The most CNN model compressing methods include training from scratch or fine-tuning steps to improve performance. In fact, these studies achieved significant space-saving without critical accuracy loss. Despite its great performance, it is not always appropriate method due to some drawbacks. The training step requires the full dataset that the model has been trained. Unfortunately, not all datasets are always accessible. Moreover, training the quantized models is time-consuming, Since even a lightweight model, training process is not optimized on some device while the original model trained with GPU acceleration. To avoid these drawbacks, in this paper, we propose compressing method that can be performed in mobile devices without the training dataset. We conducted experiments on various models and datasets to analyze whether our method could be applied to various environments.

II. Video codec based weight compression

The processes of the proposed method are as follows. A pre-trained model is quantized to 8-bits integer weights with linear symmetric quantization [1], and the fully-connected layers of the model are encoded using a High Efficiency Video Coding(HEVC) encoder [2]. In decoding processes, a bitstream of fully-connected layers is decoded with HEVC and concatenated to the whole model.

To evaluate the performance with different datasets, we compared the validation accuracy of the compressed model that trained with ImageNet [3], Places365 [4]. As a result, 76% of model size was reduced and accuracy loss was less than 3%. We also compared the performance on ResNet with other compressing methods without fine-tuning steps [5], [6]. While other methods reduced 35% of model size, we reduced 76% of size and a higher accuracy.

III. Conclusion

This paper proposes video codec based compressing method for deep neural network models. As HEVC has been optimized for a variety of mobile devices and the proposed method does not require the original dataset, this method is generally available. We have also demonstrated that this method outperforms other methods in various datasets through experiments.

Acknowledgment

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2020-0-00920, Development of ultra high resolution unstructured plenoptic video storage/compression/streaming technology for medium to large space) and MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2020-2017-0-01630) supervised by the IITP(Institute for Information & communications Technology Promotion)

References

[1] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2704–2713.

[2] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (hevc) standard," *IEEE Transactions on circuits and systems for video technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[3] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li, "Imagenet large scale visual recognition challenge," *CoRR*, vol. abs/1409.0575, 2014. [Online]. Available: http://arxiv.org/abs/1409.0575

[4] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.

[5] J.-H. Luo, J. Wu, and W. Lin, "Thinet: A filter level pruning method for deep neural network compression," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5058–5066.

[6] R. Liu, J. Cao, P. Li, W. Sun, Y. Zhang, and Y. Wang, "Nfp: A no fine-tuning pruning approach for convolutional neural network compression," in *2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD)*. IEEE, 2020, pp. 74–77.