Extended Abstract for Research Project

József Konczer
Imagination Technologies Ltd
konczer.j@gmail.com

Abstract

The proposed project is the continuation and generalization of the paper: Self-Compressing Neural Networks (1). The main contribution of the paper is a novel quantization-aware training (QAT) method. The approach can be used both to compress models during training, and to compress pretrained models. The experiments of the original paper were conducted on visual models (classification on CIFAR10 dataset), which I would extend and generalize to natural language processing models and tasks.

1 Introduction

Deep neural networks (DNNs) are a powerful tool that have shown unmatched performance in various tasks in computer vision, natural language processing and optimal control, to mention only a few. The high computational resource requirements, however, constitute one of the main drawbacks of DNNs, hindering their massive adoption on edge devices. With the growing number of tasks performed on edge devices, e.g., smartphones or embedded systems, and the availability of dedicated custom hardware for DNN inference, the subject of DNN compression has gained popularity. (2)

The objective in the paper (1) was threefold: (1) to compress networks during training to realize benefits in training time; (2) to reduce the size of weight and activation tensors by eliminating redundant channels; and (3) to reduce the number of bits required to represent weights. The second and third points produce a smaller network expected to execute more efficiently on devices supporting variable bit depth weight formats. Despite being conceptually simple, the approach was effective and the authors demonstrate high compression rates on an example classification network.

The main novelty in paper (1) was a differentiable quantization scheme, which is a representation of model weights x on a tunable but finite bit depth b and exponent of a floating-point representation e:

$$q(x, b, e) = 2^e \left\lfloor \min(\max(2^{-e}x, -2^b), 2^b - 1) \right\rceil \tag{1}$$

Where $\lfloor . \rceil$ is the rounding function which rounds to nearest integer with ties to nearest even. The quantization works for continuous b, e parameters, which allows to apply standard learning methods.

2 Related Work

Quantization-aware training (QAT) is both a relatively old concept (3), and an active research area (4). While the proposed work aims to bridge multiple research areas: low bit depth neural networks(5), QAT(6; 4; 3), and induced sparsity (particularly channel pruning)(7).

Submitted to 31st Conference on Neural Information Processing Systems (NIPS 2017). Do not distribute.

3 Visual DNN Results

The main results of the method on visual tasks are demonstrated in figures 1, 2:

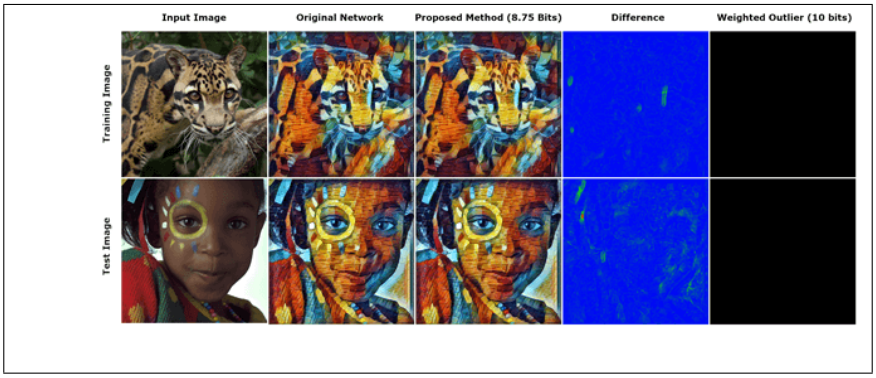
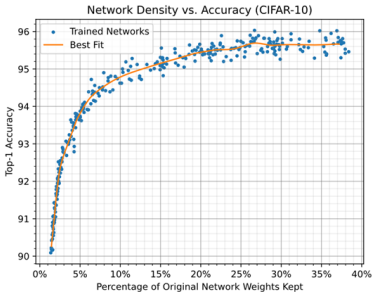
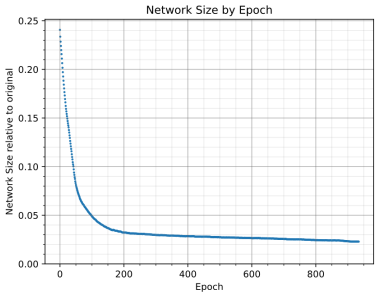


Figure 1: Style transfer results on pretrained model: the results of the original and proposed compressed model are hardly distinguishable. Source: Szabolcs Cséfalvay's blog post



(a) Accuracy-Network size relation



(b) Network size shrinking during training

Figure 2: Compression during training. Source: Szabolcs Cséfalvay's blog post

4 Language Models

The proposed self compressing method is model agnostic, therefore it could by applied to language models as well. (An example for an already existing technique is Alpha Tuning (8))

5 Outline of the Proposed Project

- Reproduction:
  - Reproducing the results on a ResNet-v2 50(9; 10)(Implemented in PyTorch)
  - trained and tested on CIFAR 10 (11)
- Extension:
  - Possibility I:
    - Self compressed BERT (12) (Implemented in PyTorch)
    - trained on Wikipedia data set e.g. WikiText-2
  - Possibility II:
    - Self compressed RoBERTa (13) (Implemented in PyTorch)
    - trained on Goodreads Dataset (14; 15)

References

[1] S. Cséfalvay and J. Imber, "Self-compressing neural networks," 2023. [Online]. Available: https://arxiv.org/abs/2301.13142

[2] Y. Nahshan, B. Chmiel, C. Baskin, E. Zheltonozhskii, R. Banner, A. M. Bronstein, and A. Mendelson, "Loss aware post-training quantization," *Machine Learning*, vol. 110, no. 11-12, pp. 3245–3262, Oct. 2021. [Online]. Available: https://doi.org/10.1007/s10994-021-06053-z

[3] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, and K. Keutzer, "A survey of quantization methods for efficient neural network inference," 2021. [Online]. Available: https://arxiv.org/abs/2103.13630

[4] A. D'efossez, Y. Adi, and G. Synnaeve, "Differentiable model compression via pseudo quantization noise," *ArXiv*, vol. abs/2104.09987, 2021.

[5] E. Wang, J. J. Davis, P. Y. K. Cheung, and G. A. Constantinides, "Lutnet: Learning fpga configurations for highly efficient neural network inference," *IEEE Transactions on Computers*, vol. 69, no. 12, pp. 1795–1808, 2020.

[6] R. Gong, X. Liu, S. Jiang, T. Li, P. Hu, J. Lin, F. Yu, and J. Yan, "Differentiable soft quantization: Bridging full-precision and low-bit neural networks," 2019. [Online]. Available: https://arxiv.org/abs/1908.05033

[7] Y. He, X. Zhang, and J. Sun, "Channel pruning for accelerating very deep neural networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[8] S. J. Kwon, J. Kim, J. Bae, K. M. Yoo, J.-H. Kim, B. Park, B. Kim, J.-W. Ha, N. Sung, and D. Lee, "Alphatuning: Quantization-aware parameter-efficient adaptation of large-scale pre-trained language models," *ArXiv*, vol. abs/2210.03858, 2022.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1512.03385

[10] ——, "Identity mappings in deep residual networks," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 630–645.

[11] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," *Computer Science Department, University of Toronto*, 2009. [Online]. Available: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

[12] ——, "Learning multiple layers of features from tiny images," *Computer Science Department, University of Toronto*, 2009. [Online]. Available: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019. [Online]. Available: https://arxiv.org/abs/1907.11692

[14] M. Wan and J. J. McAuley, "Item recommendation on monotonic behavior chains," in *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, S. Pera, M. D. Ekstrand, X. Amatriain, and J. O'Donovan, Eds. ACM, 2018, pp. 86–94. [Online]. Available: https://doi.org/10.1145/3240323.3240369

[15] M. Wan, R. Misra, N. Nakashole, and J. J. McAuley, "Fine-grained spoiler detection from large-scale review corpora," in *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, A. Korhonen, D. R. Traum, and L. Màrquez, Eds. Association for Computational Linguistics, 2019, pp. 2605–2610. [Online]. Available: https://doi.org/10.18653/v1/p19-1248