REAL END-TO-END QUANTIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

There are many recent studies addressing the congenital trade-off between parameter precision reduction and accuracy. Existing works on quantization-aware-training (QAT) all require dynamic quantization range and "fake" quantization to combat the loss of precision, which requires higher bit-width during training. This severely limited the quantized model's flexibility to be retrained/fine-tuned on 8-bit hardware. In this work, We propose a log-barrier loss and L1 loss term which could constrain and sparsify the network weight and computations outputs to ensure we would not need more bit. In the mean time, we adopt the μ-law non-linear quantization method. Our work opens up a huge potential to allow for retraining/fine tuning quantized models on 8-bit hardware. Our experiments show that our method could achieve 99.6% on MNIST dataset, 94.6% on CIFAR-10 dataset, an 71 % on ImageNet dataset.

1 INTRODUCTION

Ongoing trend to deploy deep learning models to the edge devices driven by latency and security concerns are facing significant challenges due to the growing size of state-of-the-art precision models. Recent benchmark models in language, vision and Speech tasks such as BERT Devlin et al. (2018), MaskRCNN He et al. (2017), ESPNet Karita et al. (2019) all require huge number of parameters which are prohibitive for them to be deployed off the cloud. Quantization seems to be a promising direction to downsize these models to allow for deployment on the edge, and let them run faster.

There are 2 approaches for quantization: post training and during training. Post training methods usually involve heavier penalty for accuracy since their inference graph are very different from their training graph. In order to capture the best of both (redution & accuracy), during quantization methods like quantization-aware-training (QAT) are thus developed to combat the discrepency between the training graph and inference graph. In this work, we are focusing on training-time quantization method.

Pains with current QAT
State-of-the-art quantization-aware-training methods Devlin et al. (2018); Esser et al. (2020); Rusci et al. (2020) usually involve "fake quantization", which meaning it is executed in FP32 or FP16 and its output is quantized to lower bit-width like INT8. Although doing this could help preserve precision of the trained models since it allows for dynamic ranges of each layer, it still suffer from the rescale loss and the loss caused by fewer buckets of information. As fake quantization still involves higher higher bit-width in calculation, so there is almost no speed-up benefit, and most importantly, it could not be done on lower bit-width devices. In practice, this presents a huge limitation for "fake" quantization to be deployed by lower bit-width hardware manufactures. For example, car eletronic manufacturor

change the narrative to sell more quantized training

Dynamic range, moving bitwidth prevent from real-world deployment? Real-world deployment? Types of quantization losses: fewer buckets of information, rescale loss, and inference training graph discrepency.

Motivated by the above mentioned pain of QAT, and inspired by barrier programming concept in optimization, we propose this log barrier loss and L1 loss that would produce a sparse and constrained output of the network which would enable all the parameters of the network to stay within a certain

boundry. We also encourge sparse dist of params. non-linear distribution of parameters, following the exponential distribution.

We propose a remedy to remedy the existing problem of QAT

In this paper,

1. Log barrier motivated by Lipchistzness
2. End to end quantization with global fixed-range (ensure 8 bit operation all the way, potentially use tail bound to prove stay within bound)
3. μ law quantization
4. L1 for sparsity

2 RELATED WORKS

2.1 POST TRAINING

Pioneer work by Google TensorflowJacob et al. (2018) Jain et al. (2020) Competitor works:

Cite the ICLR 2018 work WAGE Wu et al. (2018) did not have good performance on ImageNet, Cite Tensorflow LITE video's related works Reference benchmarks: Newest SOTAs: Rusci et al. (2020)

2020 ICLR New works: Stock et al. (2020) Esser et al. (2020) Uhlich et al. (2020)

Theoretical justification for quantization Hou et al. (2019)

2.2 QAT

2.3 MIX-PRECISION

Mixed-precision quantization methods quantize a neural network using multiple bit precisions across layers. Researchers found that the density of parameters at each tensor indicates a different compression rate. Thus, each tensor requires a different bit-precision.

2.4 PRUNING

Network pruning is another technique to compress a neural network. It removes the edges between neurons according to the influence of the edge on nodes until the targeted size and accuracy is obtained. After pruning, the network is retrained to compensate the loss of edges.

3 REAL QUANTIZATION

We quantize everything after each single operation, e.g. input, output, weight, gradient.

fixed range universal quantization does not need to adjust the range and precision during training.

however, we observed the range shift of parameters

non-linear quantization vs linear quantization

3.1 LOG-BARRIER

theory about log-barrier methods

3.2 SPARSITY

L_1 loss

3.3 μ-LAW QUANT

Let's denote the quantization function as $x' = q(x)$

3.4 TRICKS

3.4.1 STE

stochastic rounding

Chunk update

3.5 ADDTION, COMPOSITION, MULTIPLY AND DIVISION

4 EXPERIMENTS

4.1 MNIST

4.2 CIFAR10

4.3 IMAGENET

5 DISCUSSION

6 CONCLUSION

We achieved fixed-point quantization

REFERENCES

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Steven K Esser, Jeffrey L McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S Modha. Learned step size quantization. In *International Conference on Learning Representations*, 2020.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Lu Hou, Ruiliang Zhang, and James T Kwok. Analysis of quantized models. In *International Conference on Learning Representations*, 2019.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry 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*, pp. 2704–2713, 2018.

Sambhav Jain, Albert Gural, Michael Wu, and Chris Dick. Trained quantization thresholds for accurate and efficient fixed-point inference of deep neural networks. In I. Dhillon, D. Papailiopoulos, and V. Sze (eds.), *Proceedings of Machine Learning and Systems*, volume 2, pp. 112–128. 2020. URL `https://proceedings.mlsys.org/paper/2020/file/e2c420d928d4bf8ce0ff2ec19b371514-Paper.pdf`.

Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, et al. A comparative study on transformer vs rnn in speech applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 449–456. IEEE, 2019.

Manuele Rusci, Alessandro Capotondi, and Luca Benini. Memory-driven mixed low precision quantization for enabling deep network inference on microcontrollers. In I. Dhillon, D. Papailiopoulos, and V. Sze (eds.), *Proceedings of Machine Learning and Systems*, volume 2, pp. 326–335. 2020. URL "https://github.com/mrusci/training-mixed-precision-quantized-networks".

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

Stefan Uhlich, Lukas Mauch, Fabien Cardinaux, Kazuki Yoshiyama, Javier Alonso Garcia, Stephen Tiedemann, Thomas Kemp, and Akira Nakamura. Mixed precision dnns: All you need is a good parametrization. In *International Conference on Learning Representations*, 2020.

Shuang Wu, Guoqi Li, Feng Chen, and Luping Shi. Training and inference with integers in deep neural networks. In *International Conference on Learning Representations*, 2018.

A Appendix

You may include other additional sections here.