Optimal Quantization using Scaled Codebook

Yerlan Idelbayev[1], **Pavlo Molchanov**[2], **Maying Shen**[2], **Hongxu Yin**[2],
Miguel Á. Carreira-Perpiñán[1] and **Jose M. Alvarez**[2]
[1]Dept. CSE, University of California, Merced
[2]NVIDIA Corporation

Problem setup

Given a sorted data vector $\mathbf{w}$ with elements $w_1 \leq w_2 \leq \cdots \leq w_N$ and a fixed codebook $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$ such that $c_1 < c_2 < \cdots < c_K$ we would like to learn the optimal α-rescaling of the codebook and the assignments ($\mathbf{Z}$) of datapoints into the rescaled codebook defined by the following MSE problem:

$$\min_{\alpha,\, \mathbf{z}_1, \ldots, \mathbf{z}_N} \quad \textsc{Loss}(\alpha, \mathbf{Z}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} (w_n - \alpha\, c_k)^2 \qquad (1)$$

$$\text{s.t.} \quad \mathbf{z}_n^T \mathbf{1} = 1, \quad \mathbf{z}_n \in \{0, 1\}^K.$$

We give an $\mathcal{O}(NK \log K)$ algorithm that finds the optimal solution of this problem.

The problem definition and the algorithm is general, includes many important cases:

1. symmetric INT8 quantization: $\mathcal{C} = \{0, \pm 1, \pm 2, \ldots, \pm 127\}$
2. powers-of-two quantization: $\mathcal{C} = \{0, \pm 1, \pm 2^2, \pm 2^3, \ldots, 2^s\}$
3. log-scale codebooks: $\mathcal{C} = \{0, \pm \log 2, \pm \log 3, \pm \log 4, \ldots, \pm \log s\}$
4. and many others!

Related work

The problem we are solving is well-known, and many related work has been published.

- Previously, it was believed that problem (1) is hard to optimize [8] and requires exponential-time algorithm [13].
- Iterative or heuristic search algorithms without any optimality guarantees [1, 2, 8, 11, 13].
- Optimal algorithms for certain special cases are known: for binary codebook of $\mathcal{C} = \{-1, 1\}$ [10], for ternary codebook of $\mathcal{C} = \{-1, 0, 1\}$ [5, 12], or assuming some analytical distribution on $\mathbf{w}$ [3, 4, 6, 7, 9]

We show that we can find a globally optimal solution in polynomial time $\mathcal{O}(NK \log K)$ for any codebook $\mathcal{C}$ without any data distribution assumptions.

Overall idea of the algorithm

We build our algorithm by analyzing the properties of the optimal quantizer of (1). To do so, we analyze the loss wrt α and $\mathbf{Z}$ separately.

- **Locally optimal scale** for a fixed $\mathbf{Z}$:

$$\min_{\alpha} \textsc{Loss}(\alpha, \mathbf{Z}) \implies \alpha = \textsc{OptScale}(\mathbf{Z})$$

- **Locally optimal assignments** for a fixed α:

$$\min_{\mathbf{Z}} \textsc{Loss}(\alpha, \mathbf{Z}) \implies \mathbf{Z} = \textsc{OptAssignment}(\alpha)$$

Exact forms of $\textsc{OptScale}(\mathbf{Z})$ and $\textsc{OptAssignment}(\alpha)$ are in the paper

Overall idea of the algorithm (cont.)

Clearly, the global solution $(\alpha^*, \mathbf{Z}^*)$ must satisfy

$$\alpha^* = \textsc{OptScale}(\mathbf{Z}^*) \text{ and } \mathbf{Z}^* = \textsc{OptAssignment}(\alpha^*) \tag{2}$$

There are many pairs that satisfy the equation above, but the one obtaining the minimal loss is the $(\alpha^*, \mathbf{Z}^*)$-pair we are looking for.

Our contribution: We show that there exists at most NK pairs of $(\alpha, \mathbf{Z})$ which satisfy the fixedpoint criterion above. We give a simple enumeration algorithm which iterates over these pairs and finds globally optimal solution in $\mathcal{O}(NK \log K)$. See paper for details.

References

[1] R. Alvarez, R. Prabhavalkar, and A. Bakhtin. On the efficient representation and execution of deep acoustic models. In *Proc. of Interspeech'16*, pages 2746–2750, San Francisco, CA, Sept. 8–12 2016.

[2] S. Anwar, K. Hwang, and W. Sung. Fixed point optimization of deep convolutional neural networks for object recognition. In *Proc. of the IEEE Int. Conf. Acoustics, Speech and Sig. Proc. (ICASSP'15)*, pages 1131–1135, Brisbane, Australia, Apr. 19–24 2015.

[3] R. Banner, Y. Nahshan, and D. Soudry. Post training 4-bit quantization of convolutional networks for rapid-deployment. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NEURIPS)*, volume 32, pages 7950–7958. MIT Press, Cambridge, MA, 2019.

[4] Z. Cai, X. He, J. Sun, and N. Vasconcelos. Deep learning with low precision by half-wave Gaussian quantization. In *Proc. of the 2017 IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR'17)*, pages 5918–5926, Honolulu, HI, July 21–26 2017.

[5] M. Á. Carreira-Perpiñán and Y. Idelbayev. Model compression as constrained optimization, with application to neural nets. Part II: Quantization. arXiv:1707.04319, July 13 2017.

[6] J. Choi, S. Venkataramani, V. Srinivasan, K. Gopalakrishnan, Z. Wang, and P. Chuang. Accurate and efficient 2-bit quantized neural networks. In *Proc. of the 2nd Conf. Systems and Machine Learning (SysML 2019)*, Stanford, CA, Mar. 31 – Apr. 2 2019.

[7] J. Fang, A. Shafiee, H. Abdel-Aziz, D. Thorsley, G. Georgiadis, and J. Hassoun. Post-training piecewise linear quantization for deep neural networks. arXiv:2002.00104, Jan. 31 2020.

[8] K. Hwang and W. Sung. Fixed-point feedforward deep neural network design using weights $+1$, 0, and -1. In *2014 IEEE Workshop on Signal Processing Systems (SiPS)*, pages 1–6, Belfast, UK, Oct. 20–22 2014.

[9] D. Lin, S. Talathi, and S. Annapureddy. Fixed point quantization of deep convolutional networks. In M.-F. Balcan and K. Q. Weinberger, editors, *Proc. of the 33rd Int. Conf. Machine Learning (ICML 2016)*, pages 2849–2858, New York, NY, June 19–24 2016.

[10] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi. XNOR-net: ImageNet classification using binary convolutional neural networks. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Proc. 14th European Conf. Computer Vision (ECCV'16)*, pages 525–542, Amsterdam, The Netherlands, Oct. 11–14 2016.

[11] S. Shin, Y. Boo, and W. Sung. Fixed-point optimization of deep neural networks with adaptive step size retraining. In *Proc. of the IEEE Int. Conf. Acoustics, Speech and Sig. Proc. (ICASSP'17)*, pages 1203–1207, New Orleans, LA, Mar. 5–9 2017.

[12] P. Yin, S. Zhang, Y. Qi, and J. Xin. Quantization and training of low bit-width convolutional neural networks for object detection. arXiv:1612.06052, Aug. 17 2017.

[13] D. Zhang, J. Yang, D. Ye, and G. Hua. LQ-Nets: Learned quantization for highly accurate and compact deep neural networks. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Proc. 15th European Conf. Computer Vision (ECCV'18)*, pages 365–382, Munich, Germany, Sept. 8–14 2018.