Disentangled Loss for Low-Bit Quantization-Aware Training

Thibault Allenet[1] David Briand[1] Olivier Bichler[1] Olivier Sentieys[2]
[1]CEA-LIST, Saclay, France [2]Univ Rennes, Inria, Rennes, France
{thibault.allenet, david.briand, olivier.bichler}@cea.fr, olivier.sentieys@inria.fr

Abstract

Quantization-Aware Training (QAT) has recently showed a lot of potential for low-bit settings in the context of image classification. Approaches based on QAT are using the Cross Entropy Loss function which is the reference loss function in this domain. We investigate quantization-aware training with disentangled loss functions. We qualify a loss to disentangle as it encourages the network output space to be easily discriminated with linear functions. We introduce a new method, Disentangled Loss Quantization Aware Training, as our tool to empirically demonstrate that the quantization procedure benefits from those loss functions. Results show that the proposed method substantially reduces the loss in top-1 accuracy for low-bit quantization on CIFAR10, CIFAR100 and ImageNet. Our best result brings the top-1 Accuracy of a Resnet-18 from 63.1% to 64.0% with binary weights and 2-bit activations when trained on ImageNet.

1. Introduction

Many deep learning advances rely on increasing the number of parameters and computation power to achieve better performance. Also, the interest of deploying deep neural networks on edge mushroomed in the past few years. Critical applications with real-time constraints such as memory, latency, energy/power consumption, with specific scarce resource hardware or with privacy issues, cannot be inferred on Cloud. In this context, low-bit quantization is an elegant solution to allow significant memory footprint reduction, energy savings, and faster inference once engineered with hardware accelerators, while preserving performance and quality of results as close as possible to the floating-point reference.

The latest proposals present approaches to quantization aware training, where networks trained and quantized from scratch showed promising results for settings from 8 bits down to 2 bits [4, 9]. Those methods rely on the Cross Entropy Loss (CEL) function, *i.e.*, a combination of softmax and negative log likelihood, as it is the reference loss function for classification. A variation of the softmax was proposed by Liu *et al*. to encourage more discriminating features for image classification [13]. This research led to disruptive performance gains, especially in the face recognition domain [12,18], where the number of classes is an order of magnitude higher than academic image classification tasks. Also, Wan *et al*. used Gaussian Mixtures to formalize the classification space and encourage more discriminating features [17].

To date, the effect of those loss functions on quantization-aware training (QAT) remains unexplored. Our paper studies the quantization aware learning with disentangled loss functions for settings down to binary weights. We empirically show that training a model to output discriminative features improves its resilience to quantization. Results on CIFAR10, CIFAR100 and ImageNet datasets show the clear advantage of our approach, with significant performance gains, especially for very low-bit settings.

This paper is organized as follows. Section 2 presents some previous work on QAT as well as the foundation of disentangled loss functions. Section 3 introduces our method that takes advantage of both AMS and GML to improve the QAT procedure. Section 4 presents our experimental setup and the results obtained on relevant datasets.

2. Previous Work

To better understand the intuition behind our approach, we first give a brief review of the state-of-the-art techniques on quantization-aware training and disentangled losses.

2.1. Quantization Aware Training

Given a network $f : \mathbb{R}^n \Rightarrow \mathbb{R}$ with its parameters p, an input $x \in \mathbb{R}^n$ and its corresponding label y, we refer to quantization aware training (QAT) for classification as finding the non-differentiable quantization function q with the loss function L as

$$\min_p L[f(x, q(p)), y]. \quad (1)$$

Bengio *et al*. proposed the Straight-Through-Estimator (STE) to enable training with backpropagation [1]. The

STE method estimates the gradients of the quantized parameters assuming that the derivative of the quantization function q is the identity function. Such approximation error grows bigger as the bitwidth goes smaller hence decreasing the performance for low-bit settings. Esser *et al.* tackled this issue by scaling dynamically the gradients with a learnable step [4]. Following their method, the gradient landscape is shaped to encourage the full precision parameters towards the quantized points. Doing so, the proposed Learned Step Size Quantization (LSQ) method implicitly reduces the approximation error introduced by the STE and shows substantially better results over the previous quantization techniques. Alternatively, the Scaled Adjust Training (SAT) method introduced by Jin *et al.* directly scales the weights instead of the gradients to control the training dynamics, which yields state-of-the-art results [9]. We refer the interested readers to [9] for a detailed presentation of the quantization method.

2.2. Disentangled Losses

We qualify a loss to disentangle as it encourages the network output space to be easily discriminated with linear functions. Inspired by Large-Margin Softmax [13] and Sphereface [12], Wang *et al.* proposed an intuitive formulation of the margin softmax loss function called Additive Margin Softmax (AMS) [18]. The authors considered the propagation of features f_i (from the i-th sample with target y_i) in the linear layer without bias as scalar products for each column j of the weight matrix W. They used the geometric definition of the scalar product of Eq. (2), coupled with feature and weight normalization to rewrite the loss function applying a margin m on the target logit $W_{y_i}^T f_i$ and a scaling factor s, following Eq. (3).

$$f_i \cdot W_j = \|W_j\| \|f_i\| cos(\theta_j) \tag{2}$$

$$L_{AMS} = -\frac{1}{n} \sum_{i=0}^{n} \log \frac{e^{s \cdot (\cos \theta_{y_i} - m)}}{e^{s \cdot (\cos \theta_{y_i} - m)} + \sum_{j=1,\ j \neq y_i}^{c} e^{s \cdot (\cos \theta_j)}} \tag{3}$$

The softmax output probabilities can be interpreted as a vector of dimension n, n being the number of classes. The one-hot vectors encoding the different classes are the orthogonal vectors that construct the canonical basis of $\mathbb{R}^n$. Here, the subtracted margin m acts as a classification boundary offset, forcing the network to output features that are closer to the orthogonal vector corresponding to their label, thus reducing the intra-class variance of each class cluster in the network.

Wan *et al.* proposed to model the classification layer with Gaussian mixtures [17]. The Gaussian Mixture Loss (GML) draws the distances d_k between features f and the learned means μ_k to minimize the distance to the mean associated to the true label d_{z_i}. A positive margin factor α artificially inflates the distance d_{z_i} to help regulate the convergence of the network. Under the assumption that the covariance matrix is isotropic, the GML can be rewritten as

$$L_{GM} = -\frac{1}{n} \sum_{i=0}^{n} \log \frac{e^{-d_{z_i}(1+\alpha)}}{e^{-d_{z_i}(1+\alpha)} + \sum_{k=1,\ k \neq z_i} e^{-d_k}} \tag{4}$$

$$\text{with } d_k = \frac{1}{2}(f - \mu_k)^2 \tag{5}$$

3. Disentangled Loss Quantization Aware Training

Considering that features can be more discriminative than with CEL, we assume that low-bit quantization-aware training can benefit from a disentangled loss. Indeed, a smaller intra-class variance and a bigger inter-class difference should be more robust to the quantization noise. With CEL, the inter-class features are optimized to be orthogonal without constraint on their actual distance in the output space. While it is also true for AMS, it still allows for an additional margin on the orthogonality. On contrary, GML directly minimizes the distance between the features and their corresponding centroids, thus, minimizing the intra-class variance. The use of learned centroids instead of orthogonal features ensures that the distance between inter-class features is constrained by the distance of their respective centroids, as the features are attracted to their corresponding centroids. To reformulate, while AMS loss encourages a smaller intra-class variance than CEL, GML ensures both a smaller intra-class variance and a bigger inter-class difference than CEL. This is why our hypothesis is that there is a possibility to investigate the combination of several state-of-the-art methods: the presented disentangled loss functions with the SAT procedure [9]. In order to assess our hypothesis, we introduce Disentangled Loss Quantization Aware Training (DL-QAT), a method applying the intuitive formulation of AMS or GML loss function with the quantization-aware training method SAT [9].

4. Experiments

4.1. Training setups

All experiments use a Resnet-18 [7] with the **CIFAR10**, **CIFAR100** [10] and **ILSVRC 2012 ImageNet** dataset [3]. The batch size is 768 for CIFAR and 1024 for ImageNet. We use the same learning strategy as [9]. When training on CIFAR, the learning rates are 0.01 for SAT using CEL & DL-QAT using AMS loss and 0.2 for DL-QAT using GML. When training on ImagNet, the learning rate is 0.02 for both SAT using CEL and DL-QAT using GML. All networks are trained over 150 epochs. Finally, we use $m = 0.35$ from Eq. (3) and $\alpha = 0.7$ from Eq. (4) for CIFAR and $\alpha = 0$ for ImageNet as they give best results.

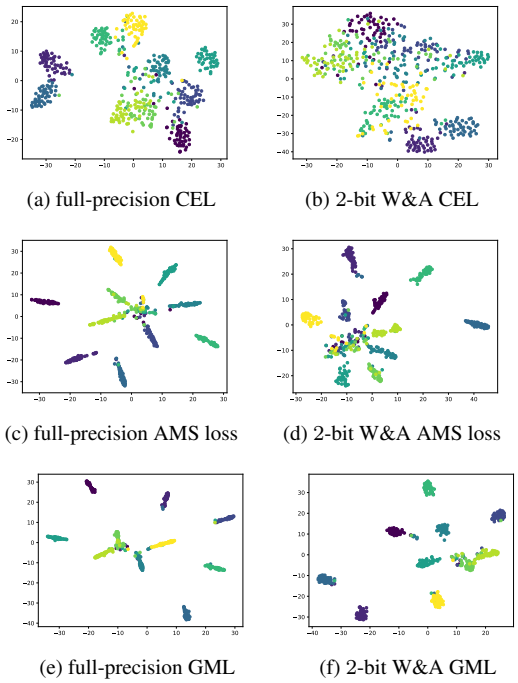
Dataset	W [bits]	A [bits]	SAT [9] Acc_{CEL}	DL-QAT (ours) Acc_{AMS}	DL-QAT (ours) Acc_{GML}	SAT [9] ΔP_{CEL}	DL-QAT (ours) ΔP_{AMS}	DL-QAT (ours) ΔP_{GML}
CIFAR10	32	32	89.4	91.7	**93.0**	–	–	–
	2	2	76.5	89.1	**91.3**	12.9	2.6	**1.7**
	binary	2	72.4	88.3	**91.2**	17.0	3.4	**1.8**
CIFAR100	32	32	66.2	68.7	**73.6**	–	–	–
	8	8	65.8	68.5	**73.1**	0.4	**0.2**	0.5
	4	4	65.4	68.4	**72.6**	0.8	**0.3**	1.0
	3	3	65.1	68.2	**73.3**	1.1	0.5	**0.3**
	2	2	61.1	66.1	**71.9**	5.1	2.6	**1.7**
	binary	8	63.9	67.9	**72.5**	2.3	**0.8**	1.1
	binary	4	63.2	67.1	**72.5**	3.0	1.6	**1.1**
	binary	3	62.4	67.0	**72.0**	3.8	1.7	**1.6**
	binary	2	59.0	65.5	**71.3**	7.2	3.2	**2.3**

Table 1. Top-1 Accuracy on vanilla Resnet-18

As it is common practice in the previous quantization approaches [4, 9], the precision of filters from the first convolution, the weights of the last layer and the activation preceding the last layer are fixed to 8 bits. Also, all batch normalization layers and the bias in the linear layer are not quantized.

4.2. Results and analysis

To better visualize the contribution of the AMS loss and the GML during quantization, we performed dimension reduction with the t-sne algorithm [16] over the input features of the linear classifier. The features fed to the t-sne algorithm are extracted from the converged Resnet-18 inferring with the same sets of 50 test images for each class. The 2D visualisations from full precision and 2-bit Resnet-18 for CEL, AMS loss and GML are plotted in Fig. 1. As expected, the full precision Resnet-18 clusters with AMS loss (c) and GML (e) are more compact than with CEL (a). It is then manifest that separating the clusters thanks to straight lines modelled by the linear classifier will be made easier. Comparing full precision in Fig. 1.(a-c-e) to 2-bit quantization in Fig. 1.(b-d-f), the clusters with the quantized version are less compact, and we can interpret this as the effect of the quantization. Comparing Fig. 1.(b) to (d) and (f), the plots show that the ambiguities caused by the quantization are reduced thanks to the disentangled losses.



(a) full-precision CEL (b) 2-bit W&A CEL

(c) full-precision AMS loss (d) 2-bit W&A AMS loss

(e) full-precision GML (f) 2-bit W&A GML

Figure 1. Dimension reduction with t-sne algorithm representing the input features of the linear classifier from CIFAR10 test data. The corresponding top-1 test accuracies are reported in Tab. 1. t-sne performed over 1000 iterations and a perplexity of 30.

4.2.1 CIFAR10 & CIFAR100

The top-1 test accuracies on CIFAR10 and CIFAR100 of the proposed DL-QAT method (Acc_{AMS}) and (Acc_{GML}) compared to the SAT method (Acc_{CEL}) are reported in Tab. 1. The lines with 32 here corresponds to single-precision floating-point, which is considered as the full precision baseline. Tab. 1 also reports the ΔP_{loss} quality metric to compare the QAT methods defined as

$$\Delta P_{loss} = Acc_{loss}^{float32} - Acc_{loss}^{quant}. \quad (6)$$

ΔP_{loss} measures the drop in top-1 accuracy between the full precision version and a quantized version of a network trained with the same loss function. Given ΔP_{GML} and ΔP_{CEL}, we can better compare the quantization resilience between disentangled losses and CEL. One noticeable result is that Resnet-18 with binary weights and 2-bit activations trained with GML (71.3%) outperforms the full precision Resnet-18 trained with CEL (66.2%). We also want to emphasize that $\Delta P_{CEL} > \Delta P_{AMS}$ and $\Delta P_{CEL} > \Delta P_{GML}$ for all settings. As the precision is reduced, the drop in top-

Method	LSQ [4]	HAWQ-V3 [20]	SAT [9]	DL-QAT GML (ours)	ABC-Net [11]	INQ [22]	SAT [9]	DL-QAT GML (ours)	LSQ [4]	SAT [9]	DL-QAT GML (ours)
W [bits]	4	4	4	4	2	2	2	2	2	2	2
A [bits]	$4/32^*$	4	4	4	32	32	8	8	$2/32^*$	2	2
Top-1 Acc	**71.1**	68.5	70.0	**70.1**	63.7	66.0	67.4	**67.9**	**67.6**	63.1	**64.0**

* For LSQ, the residual connections remain in the accumulation dynamic.
SAT and DL-QAT results are obtained from our experiments, all the other results are reported from the original papers.
DL-QAT, SAT → original Resnet-18. LSQ → full pre-activation Resnet-18.

Table 2. ImageNet Top-1 Accuracy for low-bit quantization settings of Resnet-18.

Method	BWN [15]	ABC-Net [11]	BWNH [8]	DSQ [6]	Q-Nets [19]	IR-Net [14]	SYQ [5]	PACT [2]	LQ-Net [21]	SAT [9]			DL-QAT GML (ours)		
W [bits]	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
A [bits]	32	32	32	32	32	32	8	2	2	8	4	2	8	4	2
Top-1 Acc	60.8	62.8	64.3	63.7	66.5	66.5	62.9	62.9	62.6	67.5	66.9	63.1	**67.5**	**67.2**	**64.0**

SAT and DL-QAT results are obtained from our experiments, all the other results are reported from the original papers.
DL-QAT, SAT → original Resnet-18. PACT → full pre-activation Resnet-18. LQ-Net → Resnet-18 type-A. BWN → Resnet-18 type-B.

Table 3. ImageNet Top-1 Accuracy for binary weights settings of Resnet-18.

1 accuracy grows larger. Our approach especially well limits the drop in top-1 accuracy for low-bit settings. Hence, the discriminative features, enforced by the AMS loss or the GML, enable more resilient quantization-aware training than the CEL, especially for low-bit settings. Overall, a clear tendency appears where $Acc_{CEL} < Acc_{AMS} < Acc_{GML}$. Indeed, GML minimizes the intra-class variance and constraint the distances of inter-class features while the AMS loss only minimizes the intra-class variance. Those results confirms our hypothesis on the loss function that both intra-class variance and inter-class difference need to be constraint.

4.2.2 ImageNet

In this section, we evaluate the performance of our method using the ImageNet dataset. Considering the results on CIFAR and our hypothesis on the losses, we chose to focus on the GML for ImageNet experiments. We report the top-1 test accuracy on ImageNet of our method DL-QAT using the GML and the SAT method [9] using CEL and other state-of-the-art approaches in Tab. 2 for low-bit settings and in Tab. 3 for binary weights settings.

As we read Tab. 2 and Tab. 3 from left to right, the quantization is more and more aggressive. Considering our experimental results only (DL-QAT using GML and SAT using CEL), the gap between the disentangled loss GML and the CEL is getting bigger as the settings reach more extreme quantization. Ultimately, in the binary weights and 2-bit activation setting, our approach reaches an accuracy of $64,0\%$, improving by 0.9% the CEL score of 63.1%.

When comparing our method to the other approaches, the version of Resnet-18 and the quantization method matter. Notably, the Resnet-18 results reported in Esser *et al*. [4] use pre-activation quantization scaling and thus keep the residual connections in the same precision as the accumulation (*i.e*., 32 bits). While this significantly improves the final accuracy in low precision, the actual precision of the dataflow is not strictly the activation's precision. For this reason, we have chosen to keep Resnet-18 with post-activation for our experiments, which makes it however not fully comparable with the LSQ reported results. For 2-bit weights, our method achieves substantial improvement over ABC-Net [11] and INQ [22], while the setting is more constraining on the activations. One noticeable result over the binary weights experiments is that our method with 4-bit activations reaches 67.2% and surpasses all other approaches with full precision or 8-bit activations. Looking at the stricter quantization setting with binary weights and 2-bit activations, our approach achieves the highest performance with 64% top-1 accuracy. Over all approaches, our method demonstrates the best performance on ImageNet for extreme quantization.

5. Conclusion

In this paper, we target very-low-bit settings and propose to study multiple losses to further reduce the gap in accuracy of quantization. We introduce DL-QAT, a method combining quantization-aware training and disentangled losses, as our tool to investigate the contribution of those different loss functions for extreme quantization. Preliminary experiments on CIFAR10 and CIFAR100 are conducted to visualise and lighten the advantage of our method. Further results on ImageNet show that our approach improves by nearly 1% the top-1 accuracy of Resnet-18 with binary weights and 2-bit activations. Overall, the experiments confirm our hypothesis and encourage future use and research of disentangled losses for Quantization Aware Training.

Acknowledgements

This work was supported by the French Agence Nationale de la Recherche grant AdequatDL ANR-18-CE23-0012.

References

[1] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013. 1

[2] Jungwook Choi, Zhuo Wang, Swagath Venkataramani, Pierce I-Jen Chuang, Vijayalakshmi Srinivasan, and Kailash Gopalakrishnan. Pact: Parameterized clipping activation for quantized neural networks. *arXiv preprint arXiv:1805.06085*, 2018. 4

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE CVPR*, pages 248–255, 2009. 2

[4] Steven K Esser, Jeffrey L McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S Modha. Learned step size quantization. In *ICLR*, 2020. 1, 2, 3, 4

[5] Julian Faraone, Nicholas Fraser, Michaela Blott, and Philip HW Leong. Syq: Learning symmetric quantization for efficient deep neural networks. In *IEEE CVPR*, pages 4300–4309, 2018. 4

[6] Ruihao Gong, Xianglong Liu, Shenghu Jiang, Tianxiang Li, Peng Hu, Jiazhen Lin, Fengwei Yu, and Junjie Yan. Differentiable soft quantization: Bridging full-precision and low-bit neural networks. In *IEEE ICCV*, pages 4852–4861, 2019. 4

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE CVPR*, pages 770–778, 2016. 2

[8] Qinghao Hu, Peisong Wang, and Jian Cheng. From hashing to cnns: Training binary weight networks via hashing. In *AAAI*, 2018. 4

[9] Qing Jin, Linjie Yang, and Zhenyu Liao. Towards efficient training for neural network quantization. *arXiv preprint arXiv:1912.10207*, 2019. 1, 2, 3, 4

[10] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2

[11] Xiaofan Lin, Cong Zhao, and Wei Pan. Towards accurate binary convolutional neural network. *arXiv preprint arXiv:1711.11294*, 2017. 4

[12] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *IEEE CVPR*, pages 212–220, 2017. 1, 2

[13] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In *ICML*, volume 2, page 7, 2016. 1, 2

[14] Haotong Qin, Ruihao Gong, Xianglong Liu, Mingzhu Shen, Ziran Wei, Fengwei Yu, and Jingkuan Song. Forward and backward information retention for accurate binary neural networks. In *IEEE CVPR*, pages 2250–2259, 2020. 4

[15] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. In *ECCV*, pages 525–542. Springer, 2016. 4

[16] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 3

[17] Weitao Wan, Yuanyi Zhong, Tianpeng Li, and Jiansheng Chen. Rethinking feature distribution for loss functions in image classification. In *IEEE CVPR*, pages 9117–9126, 2018. 1, 2

[18] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018. 1, 2

[19] Jiwei Yang, Xu Shen, Jun Xing, Xinmei Tian, Houqiang Li, Bing Deng, Jianqiang Huang, and Xian-sheng Hua. Quantization networks. In *IEEE CVPR*, pages 7308–7316, 2019. 4

[20] Zhewei Yao, Zhen Dong, Zhangcheng Zheng, Amir Gholami, Jiali Yu, Eric Tan, Leyuan Wang, Qijing Huang, Yida Wang, Michael Mahoney, et al. Hawq-v3: Dyadic neural network quantization. In *International Conference on Machine Learning*, pages 11875–11886. PMLR, 2021. 4

[21] Dongqing Zhang, Jiaolong Yang, Dongqiangzi Ye, and Gang Hua. Lq-nets: Learned quantization for highly accurate and compact deep neural networks. In *ECCV*, pages 365–382, 2018. 4

[22] Aojun Zhou, Anbang Yao, Yiwen Guo, Lin Xu, and Yurong Chen. Incremental network quantization: Towards lossless cnns with low-precision weights. *arXiv preprint arXiv:1702.03044*, 2017. 4