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.