IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 25 June 2023, accepted 11 July 2023, date of publication 21 July 2023, date of current version 1 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3297890

RESEARCH ARTICLE

Feature Affinity Assisted Knowledge Distillation and Quantization of Deep Neural Networks on Label-Free Data

ZHIJIAN LI[1], BIAO YANG[1], PENGHANG YIN[2], YINGYONG QI[1], AND JACK XIN[1]
[1]Department of Mathematics, University of California, Irvine, Irvine, CA 92617, USA
[2]Department of Mathematics and Statistics, State University of New York at Albany, Albany, NY 12246, USA

Corresponding author: Zhijian Li (zhijil2@uci.edu)

This work was supported in part by NSF under Grant DMS-1924935, Grant DMS-1952644, Grant DMS-2151235, and Grant DMS-2208126; and in part by the Qualcomm Faculty and Gift Awards.

ABSTRACT In this paper, we propose a feature affinity (FA) assisted knowledge distillation (KD) method to improve quantization-aware training of deep neural networks (DNN). The FA loss on intermediate feature maps of DNNs plays the role of teaching middle steps of a solution to a student instead of only giving final answers in the conventional KD where the loss acts on the network logits at the output level. Combining logit loss and FA loss, we found via convolutional network experiments on CIFAR-10/100, and Tiny ImageNet data sets that the quantized student network receives stronger supervision than from the labeled ground-truth data. The resulting FA quantization-distillation (FAQD), trained to convergence with a cosine annealing scheduler for 200 epochs, is capable of compressing models on label-free data up to or exceeding the accuracy levels of their full precision counterparts, which brings immediate practical benefits as pre-trained teacher models are readily available and unlabeled data are abundant. In contrast, data labeling is often laborious and expensive. Finally, we propose and prove error estimates for a fast feature affinity (FFA) loss function that accurately approximates FA loss at a lower order of computational complexity, which helps speed up training for high resolution image input. Source codes are available at: https://github.com/lzj994/FAQD

INDEX TERMS Model compression, quantization, knowledge distillation, image classification, convolutional neural networks.

I. INTRODUCTION

Quantization is one of the most popular methods for deep neural network compression, by projecting network weights and activation functions to lower precision thereby accelerate computation and reduce memory consumption. However, there is inevitable loss of accuracy in the low bit regime. One way to mitigate such an issue is through knowledge distillation (KD [9]). In this paper, we study a feature affinity assisted KD so that the student and teacher networks not only try to match their logits at the output level but also match feature maps in the intermediate stages. This is similar to teaching a student through intermediate steps of a solution instead of just showing the final answer (as in conventional KD [9]). Our method does not rely on ground truth labels while enhancing student network learning and closing the gaps between full and low precision models.

The associate editor coordinating the review of this manuscript and approving it for publication was Li He.

A. WEIGHT QUANTIZATION OF NEURAL NETWORK

Quantization-aware training (QAT) searches the optimal model weight in training. Given an objective L, the classical QAT scheme ([6], [20]) is formulated as

$$\begin{cases} w^{t+1} = w^t - \nabla_u L(u^t), \\ u^{t+1} = \text{Quant}(w^{t+1}), \end{cases} \tag{1}$$

where Quant is projection to a low precision quantized space. Yin et al. [27] proposed BinaryRelax, a relaxation form of

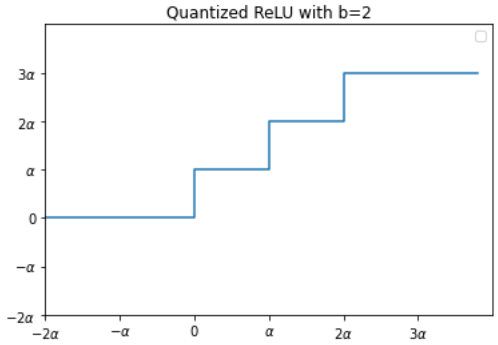


FIGURE 1. Plot of 2-bit quantized ReLU $\sigma(x, \alpha)$.

QAT, which replaces the second update of (1) by

$$u^{t+1} = \frac{w^{t+1} + \lambda^{t+1}\text{Quant}(w^{t+1})}{1 + \lambda^{t+1}},$$
$$\lambda^{t+1} = \eta\lambda^t \quad \text{with } \eta > 1. \tag{2}$$

Darkhorn et al. [7] further improved (2) by designing a more sophisticated learnable growing scheme for λ^t and adding a learnable parameter into Quant$(\cdot)$. Polino et al. [18] proposed quantized distillation (QD), a QAT framework that leverages knowledge distillation for quantization. Under QD, the quantized model receives supervision from both ground truth (GT) labels and a trained teacher in float precision (FP). The objective function has the generalized form ($\alpha \in (0, 1)$):

$$\mathcal{L}_{QD} = \alpha\mathcal{L}_{KD} + (1 - \alpha)\mathcal{L}_{GT} \tag{3}$$

where $\mathcal{L}_{KD}$ is Kullback–Leibler divergence (KL) loss, and $\mathcal{L}_{GT}$ is negative log likelihood (NLL) loss. In order to compare different methods fairly, we introduce two technical terms: end-to-end quantization and fine-tuning quantization. End-to-end quantization is to train a quantized model from scratch, and fine-tuning quantization is to train a quantized model from a pre-trained float precision (FP) model. With the same method, the latter usually lands a better result than the former. Li et al. [14] proposed a mixed quantization (a.k.a. BRECQ) that takes a pre-trained model and partially retrains the model on a small subset of data.

B. ACTIVATION QUANTIZATION

In addition to weight quantization, the inference of neural networks can be further accelerated through activation quantizaton. Given a resolution $\alpha > 0$, a quantized ReLU activation function of bit-width $b \in \mathbb{N}$ is $\sigma = \sigma(x, \alpha)$:

$$\sigma = \begin{cases} 0 & x < 0 \\ k\alpha & (k-1)\alpha \le x < k\alpha, \quad 1 \le k \le 2^b - 1 \\ (2^b - 1)\alpha & x \ge (2^b - 1)\alpha \end{cases} \tag{4}$$

where the resolution parameter α is learned from data. A plot of 2-bit quantized ReLU is shown in Fig. 1. However, such quantized activation function leads to vanished gradient during training, which makes the standard backpropagation inapplicable. Indeed, it is clear that $\frac{\partial\sigma}{\partial x} = 0$ almost everywhere. Bengio et al. [2] proposed to use a straight through estimator (STE) in backward pass to handle the zero gradient issue. The idea is to simply replace the vanished $\frac{\partial\sigma}{\partial x}$ with a non-trivial derivative $\frac{\partial\tilde{\sigma}}{\partial x}$ of a surrogate function $\tilde{\sigma}(x, \alpha)$. Theoretical studies on STE and convergence vs. recurrence issues of training algorithms have been conducted in ([16] and [26]). Among a variety of STE choices, a widely-used STE is the x-derivative of the so-called clipped ReLU [3] $\tilde{\alpha}(x, \alpha) = \min\{\max\{x, 0\}, (2^b - 1)\alpha\}$, namely,

$$\frac{\partial\tilde{\sigma}}{\partial x} = \begin{cases} 1 & 0 < x < (2^b - 1)\alpha \\ 0 & \text{else.} \end{cases}$$

In addition, a few proxies of $\frac{\partial\sigma}{\partial\alpha}$ have been proposed ([4], [28]). In this work, we follow [28] and use the three-valued proxy:

$$\frac{\partial\sigma}{\partial\alpha} \approx \begin{cases} 0 & x \le 0 \\ 2^{b-1} & 0 < x < (2^b - 1)\alpha \\ 2^b - 1 & x \ge (2^b - 1)\alpha. \end{cases} \tag{5}$$

C. KNOWLEDGE DISTILLATION

Several works have proposed to impose closeness of the probabilistic distributions between the teacher and student networks, e.g. similarity between feature maps. A flow of solution procedure (FSP) matrix in [25] measures the information exchange between two layers of a given model. Then l_2 loss regularizes the distance between FSP matrices of teacher and student in knowledge distillation. An attention transform (AT) loss [29] directly measures the distance of feature maps outputted by teacher and student, which enhances the learning of student from teacher. Similarly, feature affinity (FA) loss [23] measures the distance of two feature maps. In a dual learning framework for semantic segmentation [23], the FA loss is applied on the output feature maps of a segmentation decoder and a high-resolution decoder. In [24], FA loss is applied on multi-resolution paths in knowledge distillation of semantic segmentation models. It improves mean Average Precision of the lightweight student model. Given two feature maps with the same height and width (interpolate if different), $\mathbf{F}^S \in \mathbb{R}^{C_1\times H\times W}$ and $\mathbf{F}^T \in \mathbb{R}^{C_2\times H\times W}$, we first normalize the feature map along the channel dimension. Given a pixel of feature map $F_i \in \mathbb{R}^C$, we construct an affinity matrix $\mathbf{S} \in \mathbb{R}^{WH\times WH}$ as:

$$\mathbf{S}_{ij} = \|\boldsymbol{F}_i - \boldsymbol{F}_j\|_\theta := \cos\theta_{ij} = \frac{\langle \boldsymbol{F}_i, \boldsymbol{F}_j\rangle}{||\boldsymbol{F}_i||||\boldsymbol{F}_j||}.$$

where θ_{ij} measures the angle between F_i and F_j. Hence, the FA loss measures the similarity of pairwise angular distance between pixels of two feature maps, which can be formulated as

$$L_{fa}(\mathbf{F}^S, \mathbf{F}^T) = \frac{1}{W^2H^2}||\mathbf{S}^T - \mathbf{S}^S||_2^2. \tag{6}$$

D. CONTRIBUTIONS

In this paper, our main contributions are:

1) We find that using mean squares error (MSE) gives better performance than KL on QAT, which is a significant improvement of QD ([18]).
2) We consistently improve the accuracies of various quantized student networks by imposing the FA loss on feature maps of each convolutional block. We also unveil the theoretical underpinning of feature affinity loss in terms of the celebrated Johnson-Lindenstrass lemma for low-dimensional embeddings.
3) We achieve state-of-art quantization accuracy on CIFAR-10, CIFAR-100, and Tiny ImageNet. Our FAQD framework *can train a quantized student network on unlabeled data* up to or exceeding the accuracy of its full precision counterpart.
4) We propose a randomized Fast FA (FFA) loss to accelerate the computation of training loss, and prove its convergence and error bound.

E. ORGANIZATION

This paper is organized as follows: In Sec. II, we introduce the main objective of FAQD. In particular, we present feature affinity loss and go over the comparison between MSE and KL loss. In Sec. III, we numerically verify that FAQD outperforms baseline methods. In Sec. IV, we introduce Fast feature affinity loss and verify its acceleration to FAQD.

II. FEATURE AFFINITY ASSISTED DISTILLATION AND QUANTIZATION

A. FEATURE AFFINITY LOSS

In quantization setting, it is unreasonable to require that F^S be close to F^T, as they are typically in different spaces ($F^S \in \mathcal{Q}$ in full quantization) and of different dimensions. However, F^S can be viewed as a compression of F^T in dimension, and preserving information under such compression has been studied in compressed sensing. Researchers ([19], [21]) have proposed to compress graph embedding to lower dimension so that graph convolution can be computed efficiently. In K-means cluttering problem, several methods ([1], [17]) have been designed to project the data into a low-dimensional space such that

$$||\text{Proj}(\mathbf{x}) - \text{Proj}(\mathbf{y})|| \approx ||\mathbf{x} - \mathbf{y}||, \quad \forall\, (\mathbf{x}, \mathbf{y}), \tag{7}$$

and so pairwise distances from data points to the centroids can be computed at a lower cost.

In view of the feature maps of student model as a compression of teacher's feature maps, we impose a similar property in terms of pairwise angular distance:

$$||\boldsymbol{F}_i^S - \boldsymbol{F}_j^S||_\theta \approx ||\boldsymbol{F}_i^T - \boldsymbol{F}_j^T||_\theta, \ \forall\, (i, j)$$

which is realized by minimizing the feature affinity loss. On the other hand, a Johnson–Lindenstrauss (JL [10]) like lemma can guarantee that we have student's feature affinity matrix close to the teacher's, provided that the number of channels of student network is not too small. In contrast, the classical JL lemma states that a set of points in a high-dimensional space can be embedded into a space of much lower dimension in such a way that the *Euclidean* distances between the points are nearly preserved. To tailor it to our application, we prove the following JL-like lemma in the angular distance case:

Theorem 2.1 (JohnsonLindenstrauss lemma, Angular Case): Given any $\epsilon \in (0, 1)$, an embedding matrix $\mathbf{F} \in \mathbb{R}^{n\times d}$, for $k \in (16\epsilon^{-2} \ln n, d)$, there exists a linear map $T(\mathbf{F}) \in \mathbb{R}^{n\times k}$ so that

$$\begin{aligned}(1-\epsilon)||\mathbf{F}_i - \mathbf{F}_j||_\theta \le ||T(\mathbf{F})_i - T(\mathbf{F})_j||_\theta \\ \le (1+\epsilon)||\mathbf{F}_i - \mathbf{F}_j||_\theta, \ \ \forall\ 1 \le i, j \le n\end{aligned} \tag{8}$$

where $||\boldsymbol{F}_i - \boldsymbol{F}_j||_\theta = \frac{\langle \boldsymbol{F}_i, \boldsymbol{F}_j\rangle}{\|\boldsymbol{F}_i\|\|\boldsymbol{F}_j\|}$ is the angular distance.

It is thus possible to reduce the embedding dimension down from d to k, while roughly preserving the pairwise angular distances between the points. In a convolutional neural network, we can view intermediate feature maps as $\mathbf{F}^S \in \mathbb{R}^{HW\times C_1}$ and $\mathbf{F}^T \in \mathbb{R}^{HW\times C_2}$, and feature affinity loss will help the student learn a compressed feature embedding. The FA loss can be flexibly placed between teacher and student in different positions (encoder/decoder, residual block, etc.) for different models. In standard implementation of ResNet, residual blocks with the same number of output channels are grouped into a sequential layer. We apply FA loss to the features of such layers.

$$\mathcal{L}_{FA} = \sum_{l=1}^{L} L_{fa}(\mathbf{F}_l^T, \mathbf{F}_l^S)$$

where $\mathbf{F}_l^T$ and $\mathbf{F}_l^S$ are the feature maps of teacher and student respectively. For example, the residual network family of ResNet20, ResNet56, ResNet110, and ResNet164 have $L = 3$, whereas the family of ResNet18, ResNet34, and ResNet50 have $L = 4$.

B. CHOICE OF LOSS FUNCTIONS

In this work, we propose two sets of loss function choices for the end-to-end quantization and pretrained quantization, where end-to-end quantization refers to having an untrained student model with randomly initialized weights. We investigate both scenarios of quantization and propose two different strategies for each.

The Kullback–Leibler divergence (KL) is a metric of the similarity between two probabilistic distributions. Given a ground-truth distribution P, it computes the relative entropy of a given distribution Q from P:

$$\mathcal{L}_{KL}(P||Q) = \sum_{x\in\mathcal{X}} P(x) \ln \frac{P(x)}{Q(x)}. \tag{9}$$

While KD is usually coupled with KL loss ([9], [18]), it is not unconventional to choose other loss functions. Kim et al. [13] showed that MSE, in certain cases, can outperform KL in the classic teacher-student knowledge distillation setting. KL loss

TABLE 1. Comparision of KL loss and MSE loss on CIFAR-10 data set. All teachers are pre-trained FP models, and all students are initial models (end-to-end quantization).

Student	Teacher	1-bit	2-bit	4-bit
$\mathcal{L}_{KD}$ = KL loss in (3)				
ResNet20	ResNet110	89.06%	90.86%	92.01%
$\mathcal{L}_{KD}$ = MSE in (3)				
ResNet20	ResNet110	**90.00%**	**91.01%**	**92.17%**

is also widely used for trade-off between accuracy and robustness under adversarial attacks, which can be considered as self-knowledge distillation. Given a classifier f, an original data point $\mathbf{x}$ and its adversarial example $\mathbf{x}$', TRADES [30] is formulated as

$$L_{TRADES} = \mathcal{L}_{CE}(f(x), y) + \mathcal{L}_{KL}(f(x)||f(x')).$$

Li et al. [15] showed that $L_{CE}\big(f(\mathbf{x}'), y\big)$ outperforms $\mathcal{L}_{KL}(f(x)||f(x'))$ both experimentally and theoretically.

Inspired by the studies above, we conduct experiments on different choices of the loss function. We compare KD on quatization from scratch (end-to-end). As shown in Tab. 1, MSE outperforms KL in quantization.

On the other hand, we find that KL loss works better for fine-tuning quantization. One possible explanation is that when training from scratch, the term $\ln \frac{P(x)}{Q(x)}$ is large. However, the derivative of logarithm is small at large values, which makes it converge slower and potentially worse. On the other hand, when $\frac{P(x)}{Q(x)}$ is close to 1, the logarithm has sharp slope and converges fast.

C. FEATURE AFFINITY ASSISTED DISTILLATION AND QUANTIZATION

Inspired by previous studies ([12], [14], [18]), we propose a feature affinity assisted quantized distillation (FAQD). The end-to-end quantization objective function is formulated as:

$$\begin{aligned}\mathcal{L} &= \alpha\, \mathcal{L}_{KD} + \beta\, \mathcal{L}_{FA} + \gamma\, \mathcal{L}_{GT} \\ &= \alpha\, \mathcal{L}_{MSE}\big(f^T(\mathbf{x}), f^S(\mathbf{x})\big) + \beta \sum_{l=1}^{L} \mathcal{L}_{fa}(\mathbf{F}_l^T, \mathbf{F}_l^S) \\ &\quad + \gamma\, \mathcal{L}_{NLL}(f^S(\mathbf{x}), y). \end{aligned} \tag{10}$$

In fine-tuning quantization, we replace MSE loss in (10) by KL divergence loss. In FAQD, the student model learns not only the final logits of the teacher but also the intermediate extracted feature maps of the teacher using feature affinity norm computed as in [23].

In addition to (10), we also propose a label-free objective which does not require the knowledge of labels:

$$\mathcal{L}_{\text{label-free}} = \alpha \mathcal{L}_{MSE}\big(f^T(\mathbf{x}), f^S(\mathbf{x})\big) + \beta \sum_{l=1}^{L} \mathcal{L}_{fa}(\mathbf{F}_l^T, \mathbf{F}_l^S). \tag{11}$$

Despite the pre-trained computer vision models being available from cloud service such as AWS and image/video data abundantly collected, the data labeling is still expensive and time consuming. Therefore, a label-free quantization framework has significant value in the real world. In this work, we verify that the FA loss can significantly improve KD performance. The label-free loss in Eq. (11) can outperform the baseline methods in Tab. 2 as well as the prior supervised QD in (3).

TABLE 2. End-to-end quantization accuracies of some existing quantization-aware training methods on CIFAR-10 dataset. To stick with the original work, we apply channel-wise quantization in BRECQ, denoted by ∗. All the other methods are under layer-wise quantization.

Method	1-bit	2-bit	4-bit
Model: ResNet20			
QAT ([6], [20])	87.07%	90.26%	91.47%
BinaryRelax [27]	88.64%	90.47%	91.75%
QD [18]	89.06%	90.86%	92.01%
DSQ [8]	90.24%	91.06%	91.92%
BRECQ* [14]	N/A	88.10%	89.01%

III. EXPERIMENTAL RESULTS

In Tab. 2, we listed the performance of previous methods mentioned in the introduction section. We would like to remark that the BRECQ results are from channel-wise quantization. Namely, each channel of a convolutional layer has its own float scaler and projection map. All other results in Tab. 2 are layer-wise quantization.

All experiments reported here were conducted on a desktop with Nvidia RTX6000 8GB GPU card at UC Irvine.

A. WEIGHT QUANTIZATION

In this section we test FAQD on the dataset CIFAR-10. First, we experiment on fine-tuning quantization. The float precision (FP) ResNet110 teaches ResNet20 and ResNet56. The teacher has 93.91% accuracy, and the two pre-trained models have accuracy 92.11% and 93.31% respectively. While both SGD and Adam optimization work well on the problem, we found KL loss with Adam slightly outperform SGD in this scenario. The objective is

$$\mathcal{L} = \mathcal{L}_{KL} + \mathcal{L}_{FA}$$

for the label-free quantization. When calibrating the ground-truth label, the cross-entropy loss $\mathcal{L}_{NLL}$ is used as the supervision criterion.

For end-to-end quantization, we found that MSE loss performs better than KL loss. Adam optimization struggles to reach acceptable performance on end-to-end quantization (with either KL or MSE loss). We further test the performance of FAQD on larger dataset CIFAR-100 where an FP ResNet 164 teaches a quantized ResNet110. We report the accuracies for both label-free and label-present supervision. We evaluate FAQD on both fine-tuning quantization and end-to-end quantization. In the CIFAR-100 experiment, the teacher ResNet164 has 74.50% testing accuracy. For the pretrained FAQD, the FP student ResNet110 has 72.96% accuracy. As shown in Tab. 3 and Tab. 4, FAQD has surprisingly superior performance on CIFAR-100. The binarized student almost reaches the accuracy of FP model, and the 4-bit model surpasses the FP teacher.

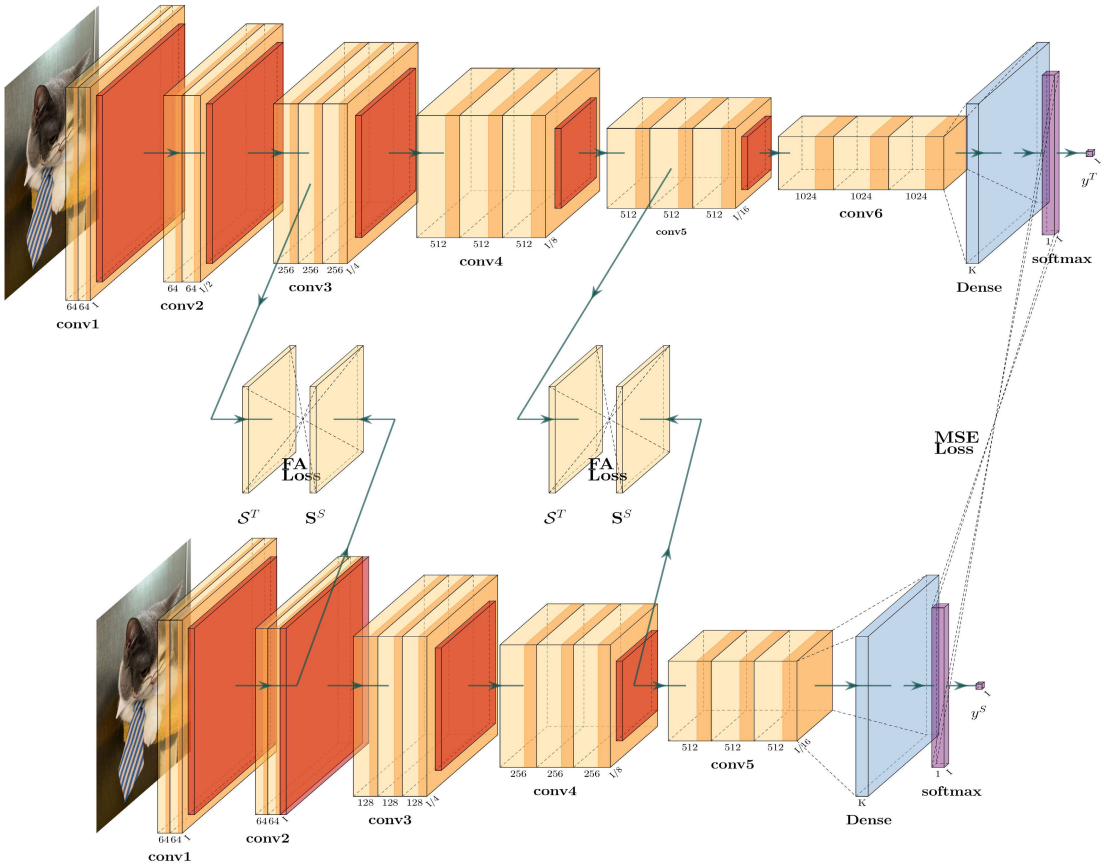


FIGURE 2. FAQD framework. The intermediate feature maps are supervised by FA loss, and the raw logits by MSE loss.

TABLE 3. Fine-tuning knowledge distillation for quantization of all convolutional layers.

Cifar-10			
Teacher ResNet110: 93.91%			
Pre-trained FP ResNet20: 92.21%			
Method	1-bit	2-bit	4-bit
Label-free FAQD	89.97%	91.40%	92.55%
FAQD with Supervision	90.92%	91.93%	92.74%
Cifar-100			
Teacher ResNet164: 74.50%			
Pre-trained FP ResNet110:72.96%			
Method	1-bit	2-bit	4-bit
label-free FAQD	73.33%	75.02%	75.78%
FAQD with Supervision	73.35%	75.24%	76.10%
Tiny ImageNet			
Teacher ResNet34: 65.60%			
Pre-trained FP ResNet18: 64.23%			
Method	1-bit	2-bit	4-bit
label-free FAQD	64.89%	65.02%	65.69%
FAQD with Supervision	65.77%	66.49%	66.62%

TABLE 4. End-to-end FAQD of ResNet110 on CIFAR-100. The accuracy of 4-bit label-free quantization surpasses 72.96% of FP ResNet110 and is close to FP ResNet164.

Cifar-10			
Teacher ResNet110: 93.91%			
Method	1-bit	2-bit	4-bit
FP student ResNet20: 92.21 %			
Label-free FAQD	89.88%	91.23%	92.19%
FAQD with Supervision	90.56%	91.65%	92.43%
Cifar-100			
Teacher ResNet164: 74.50%			
Method	1-bit	2-bit	4-bit
label-free FAQD	72.78%	74.35%	74.90%
FAQD with Supervision	73.35%	74.40%	75.31%
Tiny ImageNet			
Teacher ResNet34: 65.60%			
Method	1-bit	2-bit	4-bit
label-free FAQD	64.37%	65.05%	65.40%
FAQD with Supervision	65.13%	65.67%	65.92%

B. FULL QUANTIZATION

In this section, we extend our results to full quantization where the activation function is also quantized. In Tab. 5, we list the fine-tuning results from aforementioned methods. Among the methods in Tab. 5, only Quantized Distillation (QD) is stable under end-to-end full quantization. We extend our results to the tiny Tiny ImageNet dataset, which contains 100K downsampled 64 × 64 images across 200 classes for training. To simulate ImageNet, we interpolate the resolution back to the original 224 × 224. As shown in Tab. 6, the 4W4A fune-tuning quantization has accuracy similar to float ResNet20. Meanwhile, we close the long existing performance gap [8] when reducing activation precision to 1-bit, as the accuracy drop is linear (with respect to activation precision) and small. When fine-tuning a fully quantized

TABLE 5. Fine-tuning full quantization results of existing methods on CIFAR-10. The ∗ means the same as in Tab. 2.

Method	1W4A	4W4A
Model: ResNet20		
BinaryRelax [27]	89.22%	91.37%
QD [18]	90.15%	92.06%
BCGD [28]	89.98%	91.65%
BRECQ* [14]	N/A	88.71%

TABLE 6. End-to-end and fine-tuning full quantization on CIFAR-10, CIFAR-100 and Tiny ImageNet, with teacher networks same as in Tab. 4.

CIFAR-10			
Pretrained ResNet20: 1A32W-91.89%, 4A32W-92.01%			
Model	pre-trained	1W1A	4W4A
ResNet20	No	N/A	91.07%
ResNet20	Yes	89.70%	92.53%
CIFAR-100			
Pretrained ResNet56: 1A32W-70.96%, 4A32W-71.42%			
Model	pre-trained	1W1A	4W4A
ResNet56	No	N/A	68.84%
ResNet56	Yes	68.18	73.53%
Tiny ImageNet			
Pretrained ResNet18: 1A32W-63.82%, 4A32W-64.15%			
Model	pre-trained	1W1A	4W4A
ResNet18	No	N/A	64.67%
ResNet18	Yes	65.01	65.55%

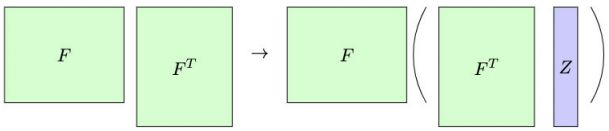


FIGURE 3. Fast feature affinity loss with a low-rank random matrix Z.

model, we follow a two-step process. First, we train an activation quantized model with floating-point weights. Subsequently, we apply full quantization using the FAQD. This technique proves to be essential, especially when scaling up the Tiny ImageNet dataset. In our experiments, we observed the following phenomenon when replacing all ReLU activation functions with 1-bit Quantized ReLU. For a pretrained 32A32W ResNet20 model, originally trained on CIFAR-10, the accuracy dropped to 80.03% from its original accuracy of 92.21%. However, when working with a pretrained ResNet-18 model on the Tiny ImageNet dataset, the accuracy plummeted to 0.62% from its initial accuracy of 64.23%.

IV. FAST FEATURE AFFINITY LOSS

A. PROPOSED METHOD

Despite the significant increase of KD performance, we note that introducing FA loss will increase the training time. If we normalize the feature maps by row beforehand, computing FA loss between multiple intermediate feature maps can be expensive.

$$\mathcal{L}_{fa}(F_1, F_2) = \|F_1F_1^T - F_2F_2^T\|_2^2. \tag{12}$$

As we freeze the pre-trained teacher, feature map of the teacher model $F_1 = f^T(\boldsymbol{x})$ is a constant, in contrast to student feature map $F_2 = f^S(\Theta, \boldsymbol{x})$. Denote $\mathbf{S}_1 = F_1F_1^T \in \mathbb{R}^{WH\times WH}$ and $g(\Theta, \boldsymbol{x}) = f^S(\Theta, \boldsymbol{x})[f^S(\Theta, \boldsymbol{x})]^T$. The feature affinity can be formulated as

$$\mathcal{L}_{fa}(\Theta) = \frac{1}{|\mathcal{X}|}\sum_{\boldsymbol{x}\in\mathcal{X}} \|\mathbf{S}_1 - g(\Theta, \boldsymbol{x})\|_2^2. \tag{13}$$

Computing $\mathbf{S}_1$ and $g(\Theta, X)$ requires $\mathcal{O}(W^2H^2C)$ complexity each (C is the number of channels), which is quite expensive. We introduce a random estimator of $L_{ffa}(\Theta)$:

$$\mathcal{L}_{ffa}(F_1, F_2, \mathbf{z}) = \frac{1}{|\mathcal{X}|}\sum_{\boldsymbol{x}\in\mathcal{X}} \|(\mathbf{S}_1 - g(\Theta, \boldsymbol{x}))\mathbf{z}\|_2^2, \tag{14}$$

where $\mathbf{z} \in \mathbb{R}^{HW}$ is a vector with i.i.d unit normal components $\mathcal{N}(0, 1)$. We show below that Eq. (14) is an unbiased estimator of FA loss (13).

Proposition 4.1:

$$\mathbb{E}_{\mathbf{z}\sim\mathcal{N}(0,1)}[\mathcal{L}_{ffa}(F_1, F_2, \mathbf{z})] = \mathcal{L}_{fa}(\Theta).$$

This estimator can achieve computing complexity $\mathcal{O}(HWC)$ by performing two matrix-vector multiplication $F_1\left(F_1^T\mathbf{z}\right)$. We define the Fast Feature Affinity (FFA) loss to be the k ensemble of (14):

$$\mathcal{L}_{ffa,k}(\Theta) = \frac{1}{|\mathcal{X}|}\sum_{\boldsymbol{x}\in\mathcal{X}} \frac{1}{k}\|(\mathbf{S}_1 - g(\Theta, \boldsymbol{x}))Z_k\|_2^2 \tag{15}$$

where $Z_k \in \mathbb{R}^{HW\times k}$ with i.i.d $\mathcal{N}(0, 1)$ components, and we have $k \ll WH$. The computational complexity of $\mathcal{L}_{ffa,k}(\Theta)$ is $\mathcal{O}(kWHC)$.

Finally, we remark that FFA loss can accelerate computation of pairwise Euclidean distance in dimensional reduction such as in (7). The popular way to compute the pairwise distance of rows for a matrix $A \in \mathbb{R}^{n\times c}$ is to broadcast the vector of row norms and compute AA^T. Given the row norm vector $v = (\|A_1\|^2, \cdots, \|A_n\|^2)$, the similarity matrix $(\mathbf{S}_{ij})$, $\mathbf{S}_{ij} = \|A_i - A_j\|^2$, is computed as

$$\mathbf{S} = \mathbf{1}\otimes v - 2AA^T + v\otimes\mathbf{1}.$$

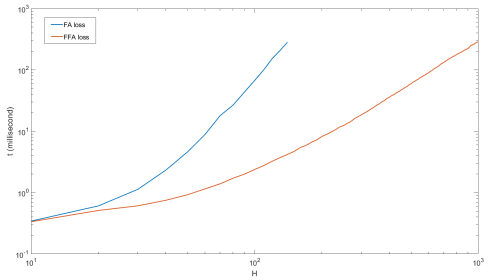
The term $2AA^T$ can be efficiently approximated by FFA loss.

B. EXPERIMENTAL RESULTS

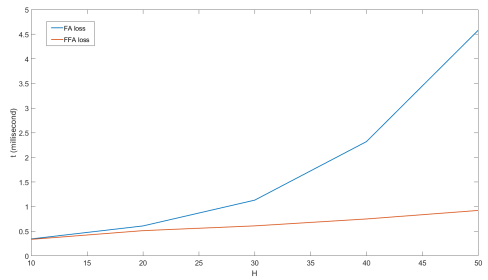
We test Fast FA loss on CIFAR-10 and Tiny ImageNet. As mentioned in the previous section, ResNet-20 has 3 residual blocks. The corresponding width and height for feature maps are 32, 16, and 8, $H = W$ for all groups, so the dimension (HW) of similarity matrices are 1024, 256, and 64. We test the fast FA loss with the number of ensemble $k = 1, 5$, and 15. The results are shown in Tab. 7. Meanwhile, FFA has added training time for each step. When $k = 1$, the accuracies are inconsistent due to large variance. With too few samples in the estimator, the fast FA norm is too noisy and jeopardizes distillation. At $k = 5$, the fast FA loss stabilizes and the accuracy improves towards that of the baseline, $\mathcal{L} = \mathcal{L}_{MSE} + \mathcal{L}_{CE}$ in Tab. 1. When k increases to 15, the performance of fast FA loss is comparable to that of the exact FA loss. Moreover, we experiment with the time

TABLE 7. 4A4W FFA accuracy and training time per epoch for ResNet20 on CIFAR-10 and ResNet18 on Tiny ImageNet, with teacher networks same as in Tab. 4. The FFA loss accelerates training and approaches the performance of exact FA loss with a proper choice of the ensemble number k.

Dataset	CIFAR-10	Tiny ImageNet
Model	ResNet20	ResNet18
Ensemble Number k	Fast FA Loss Accuracy	
1 (ResNet20)/1 (ResNet18)	88.89±2.95%	52.32± 4.35%
5/40	90.55%	56.12%
15/80	90.72%	61.12%
Ensemble Number k	Fast FA Loss Training Time Per Epoch	
1/1	29.71s	5m19s
5/40	29.77s	5m32s
15/80	30.74s	5m51s
Ensemble Number k	Exact FA Loss Training Time Per Step	
N/A	36.17s	7m36s



(a) Log-log plot for inference time of FA loss and FFA loss.



(b) Zoomed in plot for inference time of FA loss and FFA loss.

FIGURE 4. Plots for inference time of FA loss and FFA loss with $k = 1$.

consumption for computing FA loss and FFA loss. We plot the time in log scale vs. H, $(H = W)$ for feature maps. Theoretical time complexity for computing exact FA loss is $\mathcal{O}(H^4)$ and that for FFA loss is $\mathcal{O}(H^2)$. Fig. 4(a) shows the agreement with the theoretical estimate. The larger the H, the more advantageous the FFA loss. For (medical) images with resolutions in the thousands, the FFA loss will have significant computational savings. In Tab. 7, we report training time per epoch. We train models 200 epochs with cosine annealing learning rate.

C. THEORETICAL ANALYSIS OF FFA LOSS

As shown in Proposition 4.1, the FFA loss is a k-ensemble unbiased estimator of FA loss. By the strong law of large numbers, the FFA loss converges to the exact FA loss with probability 1.

Theorem 4.1: For given Θ, suppose that $|\mathcal{L}_{fa}(\Theta)| < \infty$, then

$$\forall \epsilon > 0, \exists N \;\; s.t. \;\; \forall k > N, \;\; |\mathcal{L}_{ffa,k}(\Theta) - \mathcal{L}_{fa}(\Theta)| < \epsilon.$$

Namely, the FFA loss converges to FA loss pointwise:

$$\forall\, \Theta, \;\; \lim_{k\to\infty} \mathcal{L}_{ffa,k}(\Theta) = \mathcal{L}_{fa}(\Theta).$$

We also establish the following error bound for finite k.

Proposition 4.2:

$$\mathbb{P}\big(|\mathcal{L}_{ffa,k}(\Theta) - \mathcal{L}_{fa}(\Theta)| > \epsilon\big) \le \frac{C}{\epsilon^2 k},$$

where $C \le 3\,\|\mathcal{L}_{fa}(\Theta)\|_2^4$.

Proposition 4.2 says that the probability that the FFA estimation has an error beyond a target value decays like $\mathcal{O}(\frac{1}{k})$. The analysis guarantees the accuracy of FFA loss as an efficient estimator of FA loss. Another question one might ask is whether minimizing the FFA loss is equivalent to minimizing the FA loss. Denote $\Theta^* = \arg\min L_{fa}(\Theta)$ and $\Theta_k^* = \arg\min L_{ffa,k}(\Theta)$, and assume the minimum is unique for each function. In order to substitute FA loss by FFA loss, one would hope that Θ_k^* converges to Θ^*. Unfortunately, the point-wise convergence in Theorem 4.1 is not sufficient to guarantee the convergence of the optimal points, as a counter-example can be easily constructed. In the rest of this section, we show that such convergence can be established under an additional assumption.

Theorem 4.2 (Convergence in the general case): Suppose that $\mathcal{L}_{ffa,k}(\Theta)$ converges to $\mathcal{L}_{fa}(\Theta)$ uniformly, that is

$$\forall\, \epsilon > 0, \exists N \;\; s.t. \;\; \forall\, k > N, \;\; |\mathcal{L}_{ffa,k}(\Theta) - \mathcal{L}_{fa}(\Theta)| < \epsilon$$

and $|\mathcal{L}_{fa}(\Theta)| < \infty, \;\forall \Theta$. Then

$$\lim_{k\to\infty} ||\Theta_k^* - \Theta^*||^2 = 0. \tag{16}$$

The uniform convergence assumption can be relaxed if $\mathcal{L}_{fa}$ is convex in Θ. A consequence of Theorem 4.2 is below.

Corollary 2.1 (Convergence in the convex case): Let $L_{fa} : \mathbb{R}^n \to \mathbb{R}$ be convex and L-smooth, and that $\exists$ constant $M > 0$ such that $||\Theta_k^*|| \le M$, $\forall k$. Then $\mathcal{L}_{ffa,k}$ is also convex for any k, and $\lim_{k\to\infty} ||\Theta_k^* - \Theta^*||^2 = 0$.

V. CONCLUSION

We presented FAQD, a feature assisted (FA) knowledge distillation method for quantization-aware training. It couples MSE loss with FA loss and significantly improves the accuracy of the quantized student network. FAQD applies to both weight only and full quantization, and outperforms baseline Resnets on CIFAR-10/100 and Tiny ImageNet. We also analyzed an efficient randomized approximation (FFA) to the FA loss for large dimensional feature maps, which provided theoretical foundation for FFA loss to benefit future model training on high resolution images in applications.

APPENDIX

Proof of Theorem 2.1: It suffices to prove that for any set of n unit vectors in $\mathbb{R}^d$, there is a linear map nearly preserving pairwise angular distances, because the angular distance is scale-invariant.

Let T be a linear transformation induced by a random Gaussian matrix $\frac{1}{\sqrt{k}}A \in \mathbb{R}^{k\times d}$ such that $T(\boldsymbol{F}) = \boldsymbol{F}A^T$. Define the events $\mathcal{A}_{ij}^- = \{T : (1-\epsilon)\|\boldsymbol{F}_i - \boldsymbol{F}_j\|^2 \le \|T(\boldsymbol{F})_i - T(\boldsymbol{F})_j\|^2 \le (1+\epsilon)\|\boldsymbol{F}_i - \boldsymbol{F}_j\|^2 \text{ fails}\}$ and $\mathcal{A}_{ij}^+ = \{T : (1-\epsilon)\|\boldsymbol{F}_i + \boldsymbol{F}_j\|^2 \le \|T(\boldsymbol{F})_i + T(\boldsymbol{F})_j\|^2 \le (1+\epsilon)\|\boldsymbol{F}_i + \boldsymbol{F}_j\|^2 \text{ fails}\}$.

Following the proof of the classical JL lemma in the Euclidean case [22], we have:

$$P(\mathcal{A}_{ij}^-) \le 2e^{-\frac{(\epsilon^2-\epsilon^3)k}{4}}, \quad P(\mathcal{A}_{ij}^+) \le 2e^{-\frac{(\epsilon^2-\epsilon^3)k}{4}}. \tag{17}$$

Let $\mathcal{B}_{ij} = \{T : |\boldsymbol{F}_i \cdot \boldsymbol{F}_j - T(\boldsymbol{F})_i \cdot T(\boldsymbol{F})_j| > \epsilon\}$, where $\cdot$ is the shorthand for inner product. We show that $\mathcal{B}_{ij} \subset A_{ij}^- \cup \mathcal{A}_{ij}^+$ for $\|F_i\| = \|F_j\| = 1$ by showing $\mathcal{A}_{ij}^{-C} \cap \mathcal{A}_{ij}^{+C} \subset \mathcal{B}_{ij}^C$.

If $\mathcal{A}_{ij}^{-C} \cap \mathcal{A}_{ij}^{+C}$ holds, we have

$$\begin{aligned}
&4T(\boldsymbol{F})_i \cdot T(\boldsymbol{F})_j \\
&= \|T(\boldsymbol{F})_i + T(\boldsymbol{F})_j\|^2 - \|T(\boldsymbol{F})_i - T(\boldsymbol{F})_j\|^2 \\
&\le (1+\epsilon)\|\boldsymbol{F}_i + \boldsymbol{F}_j\|^2 - (1-\epsilon)\|\boldsymbol{F}_i - \boldsymbol{F}_j\|^2 \\
&= 4\boldsymbol{F}_i \cdot \boldsymbol{F}_j + 2\epsilon(\|\boldsymbol{F}_i\|^2 + \|\boldsymbol{F}_j\|^2) \\
&= 4\boldsymbol{F}_i \cdot \boldsymbol{F}_j + 4\epsilon.
\end{aligned}$$

Therefore, $\boldsymbol{F}_i \cdot \boldsymbol{F}_j - T(\boldsymbol{F})_i \cdot T(\boldsymbol{F})_j \ge -\epsilon$. By a similar argument, we have $\boldsymbol{F}_i \cdot \boldsymbol{F}_j - T(\boldsymbol{F})_i \cdot T(\boldsymbol{F})_j \le \epsilon$. Then we have $\mathcal{A}_{ij}^{-C} \cap \mathcal{A}_{ij}^{+C} \subset \mathcal{B}_{ij}^C$, and thus

$$\mathbb{P}(\mathcal{B}_{ij}) \le \mathbb{P}(\mathcal{A}_{ij}^- \cup A_{ij}^+) \le 4\exp\{-\frac{(\epsilon^2-\epsilon^3)k}{4}\}$$

and

$$\mathbb{P}(\cup_{i<j}\mathcal{B}_{ij}) \le \sum_{i<j}\mathbb{P}(\mathcal{B}_{ij}) \le 4n^2\exp\{-\frac{(\epsilon^4-\epsilon^3)k}{4}\}.$$

This probability is less than 1 if we take $k > \frac{16\ln n}{\epsilon^2}$. Therefore, there must exist a T such that $\cap_{i<j}\mathcal{B}_{ij}^C$ holds, which completes the proof.

Proof of Proposition 4.1: Letting $N = WH$, $a_{ij} = (F_1F_1^T)_{ij}$, and $b_{ij} = (F_2F_2^T)_{ij}$ in equation (14), we have:

$$\begin{aligned}
&\mathbb{E}_z\mathcal{L}_{ffa}(F_1, F_2; 2) \\
&= \mathbb{E}_z\sum_{i=1}^N(\sum_{j=1}^N |a_{ij} - b_{ij}|z_j)^2 \\
&= \mathbb{E}_z\sum_{i=1}^N(\sum_{j=1}^N |a_{ij} - b_{ij}|^2z_j^2 + 2\sum_{j\ne k}|a_{ij} - b_{ij}||a_{ik} - b_{ik}|z_jz_k) \\
&= \mathbb{E}_z\sum_{i=1}^N\sum_{j=1}^N|a_{ij} - b_{ij}|^2z_j^2 + 2\sum_{i=1}^N\sum_{j\ne k}|a_{ij} - b_{ij}||a_{ik} - b_{ik}|z_jz_k \\
&= \sum_{i=1}^N\sum_{j=1}^N |a_{ij} - b_{ij}|^2\mathbb{E}_zz_j^2 \\
&+ 2\sum_{i=1}^N\sum_{j\ne k}|a_{ij} - b_{ij}||a_{ik} - b_{ik}|\mathbb{E}_zz_jz_k \\
&= \sum_{i=1}^N\sum_{j=1}^N|a_{ij} - b_{ij}|^2 = \mathcal{L}_{fa}(F_1, F_2; 2).
\end{aligned}$$

Proof of Theorem 4.1: Given a Gaussian matrix $Z_k = [\mathbf{z}_1, \cdots, \mathbf{z}_k] \in \mathbb{R}^{n\times k}$,

$$\mathcal{L}_{ffa,k}(\Theta) = \frac{1}{k}\sum_{l=1}^k\mathcal{L}_{ffa}(F_1, F_2, \mathbf{z}_l).$$

For any fixed Θ, $\mathcal{L}_{ffa}(F_1, F_2, \mathbf{z}_l)$, $l = 1, \cdots, k$, are i.i.d random variables. Suppose the first moment of each random variable is finite, by the strong law of large numbers, $\mathcal{L}_{ffa,k}(\Theta)$ converges to $\mathbb{E}[\mathcal{L}_{ffa}(F_1, F_2, \mathbf{z}_1)]$ almost surely. In other words, $\lim_{k\to\infty}\mathcal{L}_{ffa,k}(\Theta) = \mathcal{L}_{fa}(\Theta)$ with probability 1.

Proof of Proposition 4.2: By Chebyshev's inequality, we have

$$\begin{aligned}
&\mathbb{P}\big(\big|\mathcal{L}_{ffa,k}(\Theta) - \mathbb{E}[\mathcal{L}_{ffa,k}(\Theta)]\big| > \epsilon\big) \le \\
&\frac{\mathrm{Var}(\mathcal{L}_{ffa,k}(\Theta))}{\epsilon^2} = \frac{\mathrm{Var}(\mathcal{L}_{ffa}(F_1, F_2, \mathbf{z}_1))}{\epsilon^2 k}.
\end{aligned} \tag{18}$$

In order to estimate

$$\begin{aligned}
\mathrm{Var}(\mathcal{L}_{ffa}(F_1, F_2, \mathbf{z}_1) &= \mathbb{E}[\mathcal{L}_{ffa}^2(F_1, F_2, \mathbf{z}_1)] \\
&- \big(\mathbb{E}[\mathcal{L}_{ffa}(F_1, F_2, \mathbf{z}_1)]\big)^2,
\end{aligned} \tag{19}$$

it suffices to estimate

$$\begin{aligned}
&\mathbb{E}[\mathcal{L}_{ffa}^2(F_1, F_2, \mathbf{z}_1)] = \\
&\mathbb{E}_z\big(\sum_{i=1}^N\sum_{j=1}^N|a_{ij} - b_{ij}|^2z_j^2 + \sum_{i=1}^N\sum_{j\ne k}|a_{ij} - b_{ij}||a_{ik} - b_{ik}|z_jz_k\big)^2
\end{aligned}$$

which equals (as cross terms are zeros):

$$\begin{aligned}
&\mathbb{E}_z\big(\sum_{i=1}^N\sum_{j=1}^N|a_{ij} - b_{ij}|^2z_j^2\big)^2 \\
&+ \big(\sum_{i=1}^N\sum_{j\ne k}|a_{ij} - b_{ij}||a_{ik} - b_{ik}|z_jz_k\big)^2.
\end{aligned}$$

Direct computation yields:

$$\begin{aligned}
&\sum_{i=1}^N\sum_{j=1}^N|a_{ij} - b_{ij}|^4z_j^4 + \sum_{i=1}^N\sum_{j=1}^N\sum_{l\ne i}^N|a_{ij} - b_{ij}|^2|a_{lj} - b_{lj}|^2z_j^4 \\
&+ 2\sum_{i=1}^N\sum_{j=1}^N\sum_{l\ne j}^N|a_{ij} - b_{ij}|^2|a_{il} - b_{il}|^2z_j^2z_l^2 \\
&+ \sum_{i=1}^N\sum_{j=1}^N\sum_{k=1}^N\sum_{l\ne j}^N|a_{ij} - b_{ij}|^2|a_{kl} - b_{kl}|^2z_j^2z_l^2
\end{aligned}$$

Notice that $\mathbb{E}[z_i^4] = 3$. Taking $\mathbb{E}[\cdot]$, we derive the upper bound $3\|\mathcal{L}_{fa}\|_2^4$.

Proof of Theorem 4.2: Since $\lim_{k\to\infty} \mathcal{L}_{ffa,k}(\Theta^*) = \mathcal{L}_{fa}(\Theta^*)$, it suffices to show that

$$\lim_{k\to\infty} \inf_{\Theta} \mathcal{L}_{ffa,k}(\Theta) = \mathcal{L}_{fa}(\Theta^*).$$

Note that

$$\forall\Theta,\ \lim_{k\to\infty} \mathcal{L}_{ffa,k}(\Theta) = \mathcal{L}_{fa}(\Theta) \le \mathcal{L}_{fa}(\Theta^*).$$

Then,

$$\mathcal{L}_{fa}(\Theta^*) \ge \lim_{k\to\infty} \inf_{\Theta} L_{ffa,k}(\Theta).$$

On the other hand, for arbitrary $\epsilon > 0$, we have:

$$\exists N \ \ s.t.\ \ \forall k > N \ \ |\mathcal{L}_{ffa,k}(\Theta) - \mathcal{L}_{fa}(\Theta)| < \frac{\epsilon}{2},\ \forall\Theta$$

and there exists a sequence $\{\Theta_k\}$ s.t.

$$\mathcal{L}_{ffa,k}(\Theta_k) < \inf_{\Theta} \mathcal{L}_{ffa,k}(\Theta) + \frac{\epsilon}{2}.$$

Note that $|\mathcal{L}_{ffa,k}(\Theta_k) - \mathcal{L}_{fa}(\Theta_k)| < \frac{\epsilon}{2}$ for $k > N$, so:

$$\mathcal{L}_{fa}(\Theta^*) - \epsilon \le \mathcal{L}_{fa}(\Theta_k) - \epsilon < \inf_{\Theta} \mathcal{L}_{ffa,k}(\Theta),\quad \forall k > N.$$

Since ϵ is arbitrary, taking $k \to \infty$, we have

$$\mathcal{L}_{fa}(\Theta^*) \le \lim_{k\to\infty} \inf_{\Theta} \mathcal{L}_{ffa,k}(\Theta).$$

Proof of Corollary 4.2.1: For readability, we shorthand: $\mathcal{L}_{ffa,k} = f_k$ and $\mathcal{L}_{fa} = f$. Let

$$\mathbf{H} = \frac{\nabla^2 f}{\nabla\Theta\nabla\Theta^T} \succcurlyeq \mathbf{0} \in \mathcal{R}^{n\times n}$$

be the Hessian matrix of FA loss, which is positive semi-definite by convexity of L_{fa}. Then,

$$\frac{\nabla^2 f_k}{\nabla\Theta\nabla\Theta^T} = Z_k^T \mathbf{H} Z_k \succcurlyeq \mathbf{0} \in \mathbb{R}^{k\times k}$$

which implies the convexity of f_k for all k. Moreover, it is clear that f_k is smooth for all k since

$$\begin{aligned}\|\nabla f_k(\boldsymbol{x}) - \nabla f_k(\mathbf{y})\| &= \|Z_k(\nabla f(\boldsymbol{x}) - \nabla f(\mathbf{y}))\| \\ &\le L \cdot \|Z_k\| \cdot \|\boldsymbol{x} - \mathbf{y}\|. \end{aligned} \tag{20}$$

We note that f_k is also smooth. Although we cannot claim equi-smoothness since we cannot bound $\|Z_k\|$ uniformly in k, the above is sufficient for us to prove the desired result.

For $\forall k$, given any initial parameters Θ^0, by smoothness and convexity of f_k, it is well-known that

$$\|\Theta_k^t - \Theta_k^*\| \le \|\Theta^0 - \Theta_k^*\|$$

where Θ_k^t is the parameter we arrive after t steps of gradient descent. Hence, we can pick a compact set $\mathbf{K} = \overline{B_R(\Theta^*)}$ for R large enough such that $\{\Theta_k\}_{k=1}^{\infty} \subset \mathbf{K}$ (denote $\Theta_{\infty}^* = \Theta^*$). Now, it's suffices to prove f_k converges to f uniformly on K. In fact, f_k converges to f on any compact set. To begin with, we state a known result from functional analysis ([5], [11]):

Lemma 5.1: (Uniform boundedness and equi-Lipschitz) Let $\mathcal{F}$ be a family of convex function on $\mathbb{R}^n$ and $K \subset \mathbb{R}^n$ be a compact subset. Then, $\mathcal{F}$ is equi-bounded and equi-Lipschitz on K.

This result is established in any Banach space in [11], so it automatically holds in finite dimensional Euclidean space. By Lemma 6.1, we have that the sequence $\{f_k\}_{k=1}^{\infty}$, where $f_{\infty} = f$, is equi-Lipschitz. $\forall\ >\ 0$, $\exists\delta\ >\ 0$ s.t. $|f_k(x) - f_k(y)| < \epsilon$ for all k and $x, y \in K$ when $|x - y| < \delta$. Since $\{B(x,\delta)\}_{x\in K}$ forms an open cover for K, we have a finite sub-cover $\{B(x_j,\delta)\}_{j=1}^{m}$ of K. Since there are finitely many points x_j, there exists N_ϵ such that

$$\forall k > N_\epsilon,\ \ |f_k(x_j) - f(x_j)| < \epsilon,\ \text{for } j = 1, \cdots, m.$$

For any $x \in K$, $x \in B(x_{j^*},\delta)$ for some j^*. For all $k > N_\epsilon$, we have

$$\begin{aligned}&|f_k(x) - f(x)| \le \\ &|f_k(x) - f_k(x_{j^*})| + |f_k(x_{j^*}) - f(x_{j^*})| + |f(x_{j^*}) - f(x)| \\ &\le (2\tilde{L} + 1)\epsilon \end{aligned} \tag{21}$$

where $\tilde{L}$ is the Lipschitz constant for equi-Lipschitz family. Therefore, f_k converges to f uniformly on K.

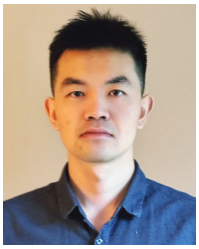
REFERENCES

[1] L. Becchetti, M. Bury, V. Cohen-Addad, F. Grandoni, and C. Schwiegelshohn, "Oblivious dimension reduction for K-means: Beyond subspaces and the Johnson-Lindenstrauss lemma," in *Proc. 51st Annu. ACM SIGACT Symp. Theory Comput.*, Jun. 2019, pp. 1039–1050.
[2] Y. Bengio, N. Léonard, and A. Courville, "Estimating or propagating gradients through stochastic neurons for conditional computation," 2013, *arXiv:1308.3432*.
[3] Z. Cai, X. He, J. Sun, and N. Vasconcelos, "Deep learning with low precision by half-wave Gaussian quantization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5406–5414.
[4] J. Choi, Z. Wang, S. Venkataramani, P. I-J. Chuang, V. Srinivasan, and K. Gopalakrishnan, "PACT: Parameterized clipping activation for quantized neural networks," 2018, *arXiv:1805.06085*.
[5] S. Cobzas, "Lipschitz properties of convex mappings," *Adv. Oper. Theory*, vol. 2, no. 1, pp. 21–49, 2017.
[6] M. Courbariaux, Y. Bengio, and J.-P. David, "BinaryConnect: Training deep neural networks with binary weights during propagations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.
[7] T. Dockhorn, Y. Yu, E. Sari, M. Zolnouri, and V. P. Nia, "Demystifying and generalizing binaryconnect," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 13202–13216.
[8] 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," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4851–4860.
[9] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.
[10] W. B. Johnson, J. Lindenstrauss, and G. Schechtman, "Extensions of Lipschitz maps into Banach spaces," *Isr. J. Math.*, vol. 54, no. 2, pp. 129–138, Jun. 1986.
[11] M. Jouak and L. Thibault, "Equicontinuity of families of convex and concave-convex operators," *Can. J. Math.*, vol. 36, no. 5, pp. 883–898, Oct. 1984.
[12] J. Kim, Y. Bhalgat, J. Lee, C. Patel, and N. Kwak, "QKD: Quantization-aware knowledge distillation," 2019, *arXiv:1911.12491*.
[13] T. Kim, J. Oh, N. Kim, S. Cho, and S.-Y. Yun, "Comparing Kullback–Leibler divergence and mean squared error loss in knowledge distillation," in *Proc. 13th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2021, pp. 2628–2635.
[14] Y. Li, R. Gong, X. Tan, Y. Yang, P. Hu, Q. Zhang, F. Yu, W. Wang, and S. Gu, "BRECQ: Pushing the limit of post-training quantization by block reconstruction," 2021, *arXiv:2102.05426*.

[15] Z. Li, B. Wang, and J. Xin, "An integrated approach to produce robust deep neural network models with high efficiency," in *Proc. Int. Conf. Mach. Learn., Optim., Data Sci.* Cham, Switzerland: Springer, 2021, pp. 451–465.
[16] Z. Long, P. Yin, and J. Xin, "Recurrence of optimum for training weight and activation quantized networks," *Appl. Comput. Harmon. Anal.*, vol. 62, pp. 41–65, Jan. 2023.
[17] K. Makarychev, Y. Makarychev, and I. Razenshteyn, "Performance of Johnson-Lindenstrauss transform for K-means and K-medians clustering," in *Proc. 51st Annu. ACM SIGACT Symp. Theory Comput.*, Jun. 2019, pp. 1027–1038.
[18] A. Polino, R. Pascanu, and D. Alistarh, "Model compression via distillation and quantization," 2018, *arXiv:1802.05668*.
[19] D. Ramasamy and U. Madhow, "Compressive spectral embedding: Sidestepping the SVD," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–15.
[20] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "XNOR-Net: ImageNet classification using binary convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 525–542.
[21] N. Tremblay, G. Puy, R. Gribonval, and P. Vandergheynst, "Compressive spectral clustering," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1002–1011.
[22] S. S. Vempala, *The Random Projection Method* (American Mathematical Society), vol. 65. Providence, RI, USA: AMS, 2005.
[23] L. Wang, D. Li, Y. Zhu, L. Tian, and Y. Shan, "Dual super-resolution learning for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3773–3782.
[24] B. Yang, F. Xue, Y. Qi, and J. Xin, "Improving efficient semantic segmentation networks by enhancing multi-scale feature representation via resolution path based knowledge distillation and pixel shuffle," in *Proc. 16th Int. Symp. Vis. Comput.* Cham, Switzerland: Springer, 2021, pp. 325–336.
[25] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7130–7138.
[26] P. Yin, J. Lyu, S. Zhang, S. Osher, Y. Qi, and J. Xin, "Understanding straight-through estimator in training activation quantized neural nets," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019, pp. 1–9.
[27] P. Yin, S. Zhang, J. Lyu, S. Osher, Y. Qi, and J. Xin, "BinaryRelax: A relaxation approach for training deep neural networks with quantized weights," *SIAM J. Imag. Sci.*, vol. 11, no. 4, pp. 2205–2223, Jan. 2018.
[28] P. Yin, S. Zhang, J. Lyu, S. Osher, Y. Qi, and J. Xin, "Blended coarse gradient descent for full quantization of deep neural networks," *Res. Math. Sci.*, vol. 6, no. 1, pp. 1–23, 2019.
[29] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," 2016, *arXiv:1612.03928*.
[30] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. E. Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7472–7482.

ZHIJIAN LI received the bachelor's degree in applied mathematics from the University of California, Los Angeles. He is currently pursuing the Ph.D. degree in mathematics with the University of California, Irvine. He held research internship positions with the Argonne National Laboratory, Didi Labs, and Uber Technologies Inc., from 2021 to 2022. He was with spatio-temporal time-series forecasting through graph-based recurrent neural networks. His research interest include the compression of neural networks, including quantization-aware training, sparsification, and channel pruning. As the leading author, he won the Springer–Verlag Best Paper Award from the 7th International Conference on Machine Learning, Optimization, and Data Science (LOD), in November 2021.

BIAO YANG received the Ph.D. degree in applied mathematics from the University of California, Irvine, in 2022. His research interests include nonconvex optimization, proximal algorithms, acceleration and compression techniques of neural networks, such as channel pruning, distillation, and neural architecture search.



PENGHANG YIN received the Ph.D. degree in applied mathematics from the University of California, Irvine, in 2016. From 2016 to 2019, he was an Assistant Adjunct Professor with the Department of Mathematics, University of California, Los Angeles. He is currently an Assistant Professor with the Department of Mathematics and Statistics, State University of New York at Albany. His research interests include applied and computational mathematics and deep learning. He won the Kovalevsky Outstanding Ph.D. Thesis Award.

YINGYONG QI received the Ph.D. degree in speech and hearing sciences from The Ohio State University, in 1989, and the Ph.D. degree in electrical and computer engineering from The University of Arizona, in 1993. He held a Faculty Member position with The University of Arizona, from 1989 to 1999. He was a Visiting Scientist with the Research Laboratory of Electronics, Massachusetts Institute of Technology, from 1995 to 1996, and a Visiting Scientist with the Visual Computing Laboratory, Hewlett Packard, Palo Alto, in 1998. He is currently the Senior Director of Technology, Qualcomm, and a Researcher with the Department of Mathematics, University of California, Irvine.

JACK XIN received the Ph.D. degree in mathematics from the Courant Institute of Mathematical Sciences, New York University, in 1990. He was a Faculty Member with The University of Arizona, from 1991 to 1999, and with The University of Texas at Austin, from 1999 to 2005. He is currently a Chancellors Professor of mathematics with UC Irvine. His research interests include applied analysis and computational methods, and their applications in multi-scale problems and data science. He is a fellow of the Guggenheim Foundation, American Mathematical Society, American Association for the Advancement of Science, and the Society for Industrial and Applied Mathematics. He was a recipient of the Qualcomm Faculty Award, in 2019 and 2022 and Qualcomm Gift Award, in 2023.