Data-Augmented Quantization-Aware Knowledge Distillation

Justin Kur, Jon Maravilla, Kaiqi Zhao
Computer Science and Engineering Department, Oakland University, Rochester, MI, USA
Email: {jakur, imaravilla, kaiqizhao}@oakland.edu

***Abstract*—Quantization-aware training (QAT) and Knowledge Distillation (KD) are combined to achieve competitive performance in creating low-bit deep learning models. Existing KD and QAT works focus on improving the accuracy of quantized models from the network output perspective by designing better KD loss functions or optimizing QAT's forward and backward propagation. However, limited attention has been given to understanding the impact of input transformations, such as data augmentation (DA). The relationship between quantization-aware KD and DA remains unexplored. In this paper, we address the question: how to select a good DA in quantization-aware KD, especially for models with low precision? We propose a novel metric which evaluates DAs according to their capacity to maximize the Generalized Contextual Mutual Information–the information not directly related to an image's labels–while also ensuring the predictions for each class are close to the ground truth labels on average. The proposed method automatically ranks and selects DAs, requiring minimal training overhead, and it is compatible with any KD or QAT algorithm. Extensive evaluations demonstrate that selecting DA strategies using our metric significantly improves state-of-the-art QAT and KD works across various model architectures and datasets.**

I. Introduction

Deep neural networks have shown outstanding performance in many areas of computer vision [1], [2]. However, these networks contain many millions of parameters and require massive amounts of computation resources, which limits their usability in resource-constrained systems and edge devices [3]. Quantization is one of the important model compression approaches to address this challenge by converting the full-precision model weights or activations to lower precision. In particular, Quantization-Aware Training (QAT) has achieved promising performance by simulating the effects of quantization during training and estimating the gradient through the quantized functions on the backward pass [4]. Recent works [5]–[8] show that the combination of QAT and knowledge distillation (KD) can improve the performance of the quantized models. However, the accuracy loss is often significant when the models are quantized to low precision [6].

Most existing works on KD and QAT focus on improving the accuracy of quantized models by refining KD loss functions or optimizing QAT's forward and backward propagation, with limited attention to the role of input transformations. Wang et al. explored the interplay between KD and data augmentation (DA) [9]. However, their analysis primarily relies on a relatively simple metric that predicts the relationship between DA and model loss instead of accuracy, and it fails to effectively estimate the quantized model's accuracy (see Sec. IV-A for details). Moreover, their study evaluates individual DA methods in isolation, which is impractical for real-world applications where augmentations are typically used in combination. The relationship between quantization-aware KD and DA remains underexplored.

In this work, we propose Data-Augmented Quantization-Aware Knowledge Distillation, an automated method for estimating the effectiveness of DA schemes for training low-precision models. Specifically, we construct a data augmentation ranking metric that maximizes Generalized Contextual Mutual Information (GCMI), the information present in an image that is not directly related to its underlying class labels, while also minimizing the deviation between the average predictions for a class and its ground truth label.

The proposed metric has several advantages. First, our metric can be computed cheaply, requiring only two passes through the training data, and no backward gradient computations or modifications to the full precision teacher model. This allows a large range of candidate DAs to be evaluated, and avoids a costly grid search as would often be performed when treating the augmentation choice as a hyperparameter. Second, because the metric is computed using only the full precision teacher model, it is inherently agnostic to the bit-width of the quantized student model. Third, the metric is capable of evaluating arbitrarily complex data augmentations. Because all that is needed to evaluate a DA is the full precision model's predictions and the ground truth image labels, any image transformation is admissible. For this work, we focus on well known DA policies, meaning we can quickly evaluate their transferability to new datasets in the context of QAT.

Our comprehensive evaluation shows that the proposed method improves the-state-of-art (SOTA) KD and QAT benchmarks under a variety of model architectures on CIFAR-10, CIFAR-100, and Tiny-ImageNet. We evaluate 7 well-known DA policies and find that the chosen DA with our metric improves top-1 test accuracy of existing SOTA QAT methods, e.g., EWGS, PACT, LSQ, and DoReFa, by up to 10% under diverse quantization levels. Furthermore, our DA improves Top-1 test accuracy across a variety of KD algorithms, improving performance on some algorithms like CRD [10] by up to 3% with identical training hyperparameters. We achieve superior Top-1 test accuracy on multiple datasets compared to existing approaches that integrate KD with QAT, including QKD [6], SPEQ [7], and SQAKD [5].

Our main contributions can be summarized as: 1) We propose a novel metric for predicting the relative performance of a quantized model trained using DAs. Our proposed metric outperforms existing baselines in DA search for KD, and it is inexpensive to compute, so it can be effectively used to reduce the search space of DAs. 2) We demonstrate training using our selected DA outperforms existing SOTA methods in QAT and QAT-aware KD. Our code is available at https://github.com/Jakur/ijcnn26.

II. Related Work

Knowledge Distillation (KD). KD transfers knowledge from large networks (termed "teacher") to improve the performance of small networks (termed "student"). KD was originally introduced by Hinton et al. [11] as an efficient method for encoding the knowledge of an ensemble into a single model by using the logits of the ensemble teacher as the target for the student. CRD [10] and RKD [12] instead attempt to match the structure of the representations at the last layer of the teacher and student. Other methods try to match the representations at the inner layers of the networks such as Attention Transfer (AT) [13] and Neuron Selectivity Transfer (NST) [14]. Finally, there are methods such as Decoupled Knowledge Distillation [15] and Refined Logit Distillation (RLD) [16] which treat the true label and other class predictions separately. In contrast to these approaches, we do not seek to modify the underlying algorithms of KD, but rather integrate KD into model quantization, so our method is orthogonal to these works.

Quantization Aware Training. Network quantization can be divided into two categories: post training quantization (PTQ) and QAT. PTQ methods do not retrain or fine tune the model, instead quantizing the weights directly to minimize the degradation from the loss in precision. In contrast, in QAT quantization is performed alongside the neural network training. QAT is more expensive than post training quantization, but in return it generally incurs a smaller penalty to model accuracy [17]. In this work, we focus on improving the accuracy of QAT-based methods.

Zhou et al. [18] introduce DoReFa-Net which expands upon initial research on binary neural networks [19] [20]. Later works introduce more learnable parameters for quantization, including activation clipping parameters in PACT [21] and LSQ [22], and parameters that control the interval between quantization levels in EWGS [23] and QIL [24]. These QAT methods have been shown to reduce the accuracy loss from quantization, but the gap between the quantized and the full precision models can still be large, especially when models are quantized to less than 8 bits. Our work aims to improve QAT model accuracy using KD and DAs.

Data Augmentation Search. Data augmentations are widely used when training deep neural networks on vision tasks, but automatically determining the best data augmentation for a given dataset or model is challenging and often computationally intensive. AutoAugment [25] uses reinforcement learning to find sets of augmentations that perform well on CIFAR, ImageNet, and SVHN. Fast AutoAugment [26] avoids the full training process and instead performs density matching between the augmented and unaugmented images. RandAugment [27] reduces the search space of augmentations by applying any augmentation from the set with uniform probability, and learns a single strength value that applies to all data augmentations. TrivialAugment [28] removes the search parameters altogether, instead opting to uniformly sample an augmentation from its set, and uniformly sample the strength with which the augmentation should be applied. Nonetheless, it can improve performance on standard models and datasets over augmentations with thousands of GPU-hours of search. AugMix [29] searches for augmentations to try to overcome dataset drift between the training and test sets. Unlike these works, we do not focus on building a new data augmentation policy from scratch, but we do look to prune the search space of existing augmentations *without* training the student model. Furthermore, we specifically define our DA evaluation metric for the case of KD.

There is also a smaller body of work focusing on data augmentations for KD. Wang et al. [30] provide a statistical justification for using the DA that minimizes the variance of the teacher's predictions. Accordingly, we benchmark our DA selection for quantization aware KD against this methodology, and find that our method is superior for models quantized to low bit precision, as shown in Sec. IV-A.

Knowledge Distillation in QAT. Recent works have integrated KD into quantized model training, but the loss in accuracy at low precision is still a significant limitation. QKD [6] proposes a three phase curriculum where first the quantized model is trained without KD, then the teacher and student learn together, and finally normal KD between the student and teacher occurs. SPEQ [7] uses a teacher whose activations are stochastically quantized to different bit levels, treating the teacher as an ensemble. CMT-KD [31] optimizes several quantized teachers and students collaboratively, performing distillation at both the logit level and in intermediate network layers. PTG [8] progressively quantizes its models into lower bit precisions, and concurrently trains the teacher alongside the student. SQAKD [5] initializes a quantized student model to its teachers weights, and performs self-supervised KD without using labels. These approaches have combined KD and QAT, but they have not considered the effects of DA at distillation time.

III. Methodology

A. Preliminaries: Quantization-Aware Training with Knowledge Distillation

Neural network quantization reduces model size and computational cost by representing weights and activations with low-bit precision (*e.g.*, 8-bit or 4-bit integers) instead of 32-bit floating-point numbers. Quantization-Aware Training (QAT) has emerged as the leading approach for obtaining high-accuracy quantized models [32], where quantization operations are simulated during training to allow the model to adapt to the reduced precision. However, QAT alone often suffers from significant accuracy degradation, particularly at aggressive quantization levels (*e.g.*, 4-bit or lower). To mitigate this performance loss, recent work has successfully

combined QAT with Knowledge Distillation (KD), where a full-precision teacher model f_t guides the training of the low-precision student f_s. In this KD-applied QAT framework, the student is trained to minimize a combined loss: $\mathcal{L}_{\text{total}} = (1-\alpha) \cdot \mathcal{L}_{\text{CE}}(y, f_s(x)) + \alpha \cdot \mathcal{L}_{\text{KD}}(f_t(x), f_s(x))$, where y denotes the ground truth label, $\alpha \in [0, 1]$ is a weighting factor, $\mathcal{L}_{\text{CE}}$ is the standard cross-entropy loss, and $\mathcal{L}_{\text{KD}}$ is the distillation loss that encourages the student's output probability distribution to match the teacher's predictions.

A central challenge is quantifying the quality of knowledge that a full-precision teacher can impart to the quantized student. While validation accuracy serves as a natural metric for teacher performance, recent work shows that the highest-accuracy model does not necessarily make the best teacher [33]. This observation motivates the development of heuristics specific to knowledge distillation. Furthermore, the choice of data augmentation can significantly impact the quality of transferred knowledge. Different augmentations expose different aspects of the teacher's learned representations, but no principled method exists for selecting augmentations that maximize the informativeness of teacher predictions while maintaining alignment with the original labels.

Problem Statement. Given a fixed full-precision teacher model f_t and a set of candidate data augmentations $\mathcal{A} = \{DA_1, DA_2, \ldots, DA_K\}$, how can we systematically select the augmentation DA^* that maximizes the knowledge transfer to a quantized student during QAT? We approach this problem by proposing a novel metric that expands upon prior work on the theoretical basis of knowledge distillation.

B. Generalized Contextual Mutual Information for Teacher Evaluation

We propose Generalized Contextual Mutual Information (GCMI) as a metric to quantify the "dark knowledge" [11] that a teacher can provide beyond simple label information. The key insight is that an effective teacher should provide instance-specific predictions that go beyond merely reproducing the average characteristics of each class.

Definition. For a teacher model f_t with output probabilities P_x for input x, we define GCMI as:

$$GCMI_{emp}(DA) = \frac{1}{N} \sum_{x_i \in D} KL(P_{x_i} \parallel Q^i_{emp}) \tag{1}$$

$$\text{where } Q^i_{emp} = \sum_{j \in [C]} P_{x_{ij}} \cdot \frac{Z_j}{||Z_j||}, Z_j = \sum_{x_k \in D} w_{j_k} P_{x_k} \tag{2}$$

where x_i are samples from dataset D, P_x is the teacher output probabilities under data augmentation DA, $P_x = f_t(DA(x))$, and w_{j_k} represents the label weight corresponding to class j for image x_k under the given augmentation. GCMI measures the KL divergence between the teacher's output probabilities P_{x_i} and the expected value of those probabilities Q^i_{emp} given the ground truth target w_{j_k}. Although Q^i_{emp} will have converged towards w_{j_k} during the teacher's training, the subtle differences in these two probability distributions constitute part of the teacher's "dark knowledge" about high level features shared between differing classes.

Intuitive Interpretation. GCMI itself quantifies, on average, how unique the predictions are for any augmented view of an image. A teacher model that provides large GCMI targets for its student presents knowledge beyond what was already present in the ground truth labels.

Connection to Prior Work. The proposed GCMI generalizes the Contextual Mutual Information (CMI) metric introduced by Ye *et al.* [34] for single-label classification. CMI can be viewed as a special case of our GCMI formulation where $w_{jk} \in \{0, 1\}$. CMI captures the aspects of one input image which are unique relative to its ground truth class. Namely, CMI is defined as: $CMI_{emp}(f) = \frac{1}{N} \sum_{y \in [C]} \sum_{x_j \in D_y} KL(P_{x_j} \parallel Q^y)$, where $Q^y = \frac{1}{|D_y|} \sum_{x_j \in D_y} P_{x_j}$, for $y \in [C]$. High CMI indicates that on average the predictions for images of a given class have high diversity.

There are two major limitations of existing CMI formulation in the context of QAT-aware KD. Firstly, CMI can only work in the case of single-label multiclass classification, which excludes a broad class of augmentations such as CutMix [35] and MixUp [36] that combine multiple images into one. For instance, in CutMix, two images A and B are combined together and the resulting ground truth of the training example is set to $y = \lambda y_A + (1-\lambda) y_B$ [35], with λ determined by the proportion of the training image originating from image A. When images A and B correspond to different classes, their CMI_{emp} would be undefined, but their $GCMI_{emp}$ can still be calculated. The second limitation is that in existing work CMI is exclusively used as an auxiliary loss term when fine-tuning a teacher model. This process is computationally expensive, especially if a separate fine-tuned model is necessary for every DA under consideration, and it causes a drop in the accuracy of the teacher model [34]. Our proposed metric addresses both of these limitations.

C. The GCMI-DEV Metric for Quantitatively Selecting Data Augmentations in KD

Rather than fine-tuning a teacher model with GCMI incorporated into the loss term, we note that it is equivalent to hold the teacher model fixed and instead modify the data to induce predictions with higher GCMI values. Concretely, this can be achieved by applying data augmentations to the underlying dataset D at distillation time. In this way, we eliminate the need for an additional model training procedure, and we avoid the decline in teacher validation accuracy that results from incorporating a secondary term into the loss function.

However, any metric which incorporates GCMI must also include a factor that rewards the teacher model's faithfulness to the ground truth labels, because GCMI alone only considers the *diversity* and not the *accuracy* of predictions. This can be achieved by ensuring that the average prediction for any given class should not deviate too much from its ground truth value (a one hot encoding of the label). The average prediction for a given class y is exactly the term Q^y, so we can quantify the

Algorithm 1 Quantization-Aware Knowledge Distillation Data Augmentation Selection

Input: Training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, candidate augmentations $\mathcal{A}$, classes C, full precision model f_T
Let $\mathcal{D}_y = \{x_i : y_i = y\}$ denote samples of class y
Output: Selected Data Augmentation DA^*

for each $DA \in \mathcal{A}$ **do**
  First pass: apply DA and compute class prototypes
  Initialize $Z_j \leftarrow \mathbf{0}$ for all $j \in [C]$
  for each $(x_k, y_k) \in \mathcal{D}$ **do**
    $x_k^{aug}, w_k \leftarrow DA(x_k, y_k; \mathcal{D})$ ▷ $w_k \in \Delta^{C-1}$: label weights
    $P_k \leftarrow f_T(x_k^{aug})$
    for each class $j \in [C]$ **do**
      $Z_j \leftarrow Z_j + w_{kj} \cdot P_k$
    end for
  end for
  $Q^y \leftarrow \text{VectorNorm}(Z_y)\ \forall y \in [C]$ ▷ $\frac{Z_y}{\|Z_y\|}$
  Second pass: compute GCMI and DEV
  $GCMI \leftarrow 0$, $DEV \leftarrow 0$
  for each $(x_i, y_i) \in \mathcal{D}$ **do**
    $x_i^{aug}, w_i \leftarrow DA(x_i, y_i; \mathcal{D})$
    $Q_{emp}^i \leftarrow \text{VectorNorm}\big(\sum_{j \in [C]} w_{ij} \cdot Q^j\big)$ ▷ Eq. (2)
    $GCMI \leftarrow GCMI + KL\big(f_T(x_i^{aug}) \parallel Q_{emp}^i\big)$ ▷ Eq. (1)
  end for
  for each class $y \in [C]$ **do**
    $DEV \leftarrow DEV + KL\big(\mathbf{1}_y \parallel Q^y\big)$ ▷ Eq. (3)
  end for
  $DEV \leftarrow DEV/C$, $GCMI \leftarrow GCMI/N$
end for
$GCMI \leftarrow \text{Normalize}(GCMI)$, $DEV \leftarrow \text{Normalize}(DEV)$
$M(DA) \leftarrow GCMI - DEV$ ▷ Eq. (4)
$DA^* \leftarrow \arg\max_{DA \in \mathcal{A}} M(DA)$

difference in the following way using their **centroid deviation** (DEV):

$$DEV(DA) = \frac{1}{C} \sum_{y \in [C]} KL(1_y \parallel Q^y) \tag{3}$$

where 1_y indicates the one-hot encoded vector corresponding to class y. Our metric aims to maximize the information gain from GCMI, while minimizing the relative entropy between the average class logits and the ground truth labels. We first normalize the GCMI and DEV according to the mean and standard deviation observed across all DAs for a model and dataset, then combine the two terms into the metric $M(DA)$:

$$M(DA) = GCMI_{normal}(DA) - DEV_{normal}(DA) \tag{4}$$

After computing $M(DA)$ for all DAs under consideration, we select the DA with the highest $M(DA)$ value. This corresponds to the DA which we predict will induce high GCMI predictions from the teacher model, without significantly harming the accuracy of the predictions from the data distribution shift. Once a DA is selected, the full precision model will be used as the teacher in the quantized model training. This entire DA selection process is outlined in Algorithm 1.

IV. Evaluation

Datasets and models. We evaluate our methodology across a range of datasets and model architectures. CIFAR-10 [37] is a standard computer vision benchmark consisting of 50,000 training images and 10,000 test images of size 32×32 with 10 total classes. On CIFAR-10, we evaluate our performance on VGG-8 and ResNet-20. CIFAR-100 [37] has the same number of images and image characteristics as CIFAR-10, but with 100 total classes. We train deeper models as compared to CIFAR-10, using VGG-13 and ResNet-32 as representative members of these two architectures. Finally, Tiny-ImageNet [38] consists of a subset of 100,000 training images from the ImageNet Large Scale Visual Recognition Challenge [39]. It contains 200 classes, and downsamples the original images to 64×64. Tiny-ImageNet presents a more computationally challenging task than CIFAR, while remaining smaller than the full ImageNet-1K dataset. We utilize both the MobileNetV2 and ResNet-18 architectures to benchmark our performance on Tiny-ImageNet.

Baselines. We compare our results against established baselines in the following four categories:

- **Quantization-aware Training (QAT)**: PACT [21], DoReFa [18], LSQ [22], EWGS [23].
- **Knowledge Distillation (KD)**: AT [13], CC [40], CRD [10], NST [14], RKD [12], RLD [16], SP [41].
- **KD + QAT**: QKD [6], SPEQ [7], SQAKD [5].
- **Data Augmentation (DA) Search in KD**: Minimizing the variance of the teacher's predictions [30].

For QAT, we compare our accuracy results to those reported from using QAT alone. For KD, we consider the quantized accuracy resulting from our selected DA against the accuracy using standard practice DAs. For KD + QAT, we compare against cited results with identical model architectures and datasets. Finally, for DA search we consider the quantized accuracy for the highest scoring DA by each metric.

Data augmentation schemes. We focus primarily on data augmentations found through automatic search, rather than hand-crafted augmentations. Specifically, we investigate the following DA policies: Auto Augment (AA) [25] which we subdivide into its 3 main policies **AA CIFAR**, **AA ImageNet**, and **AA SVHN**; **RandAugment (Rand)** [27]; **Trivial Augment (TA)** [28]; **AugMix** [29]; **CutMix** [35]. Additionally, we study **Standard DA** of random crop, random flip, and color jitter for Tiny ImageNet only, which corresponds to the lightweight DA used by existing works in QAT + KD [5].

Implementation details. Experiments are carried out on Python 3.11 using PyTorch 2.1 running on RTX 3000 series GPUs. Full precision models on CIFAR are trained using the Standard DA and CutMix. Full precision models on Tiny-ImageNet use Standard DA alone.

All CIFAR experiments use the EWGS [23] quantizer with the weight hyperparameters for labels and KD set to 1.0 and 2.0 respectively. For CIFAR-10, all models are trained with the ADAM optimizer with a learning rate of 1e-3, with a cosine learning rate decay, and a batch size of 256. EWGS has its own separate optimizer, initialized with a learning rate of 1e-5. The model's weight decay is set to 1e-4 and models are trained for 400 epochs on VGG-8, or 1200 epochs on ResNet-20 for consistency with SQAKD [5]. For CIFAR-100 all learning rates are halved, the batch size is 64, with 200 training epochs.

TABLE I: Top-1 test accuracy (%) averaged across 3 seeds for each method's DA with the standard deviation reported across the 3 trials. The baseline method selects the DA which minimizes the variance of teacher model predictions. The value in parentheses shows the difference relative to the oracle DA selection, which uses hindsight to select the best performing model (scores closer to zero are better). "$W*A\times$" denotes that the weights and activations are quantized into $*$bit and $\times$bit, respectively.

Dataset	CIFAR-10		CIFAR-100		Tiny-ImageNet	
Model	VGG-8 W2A2	ResNet-20 W2A2	VGG-13 W3A3	ResNet-32 W4A4	MobileNetV2 W3A3	ResNet-18 W3A3
Baseline	AA SVHN 92.24±0.03 (-0.60)	AA SVHN 91.28±0.08 (-0.72)	AA ImageNet 76.54±0.15 (-1.24)	AA ImageNet **75.27**±0.03 (-0.09)	CutMix 52.36±0.27 (-3.68)	CutMix 67.14±0.05 (-0.35)
Ours	TrivialAugment **92.64**±0.14 (-0.19)	RandAugment **91.82**±0.18 (-0.19)	TrivialAugment **77.77**±0.27 (-0.00)	AA ImageNet **75.27**±0.03 (-0.09)	TrivialAugment **55.28**±0.11 (-0.87)	TrivialAugment **67.28**±0.25 (-0.21)

On Tiny ImageNet, we quantize models using DoReFa, PACT, and LSQ via MQBench [42], a commonly used quantization library. Training uses SGD with an initial learning rate of 4e-3, 2500 warmup steps, and cosine decay. Momentum is set to 0.9 with weight decay 1e-4, batch size 64, and 100 training epochs. For ResNet-18, label and KD loss weights are set to 1.0 and 2.0, respectively; for MobileNetV2, they are increased to 3.0 and 6.0 while keeping the same ratio.

Edge device latency measurements are conducted on an AMD Ryzen AI 7 350 processor in single-threaded configuration. PyTorch's benchmarking utility with blocked autorange and a minimum runtime of 2 seconds is used to report a median latency over multiple runs.

A. Improving DA Selection for KD-applied QAT

Tab. I shows our augmentation search results across CIFAR-10, CIFAR-100 and Tiny-ImageNet, compared against the existing baseline [30] which selects DAs according the variance of the teacher model's predicted probabilities. We show both the DA method which is selected among the 8 possible choices, and the true accuracy of the low precision model when trained using QAT and KD. Accuracy is reported as an average of 3 experimental runs to reduce the potential impact of outliers. In the case of CIFAR-100 ResNet-32, both our metric and the baseline select the same DA method, leading to identical performance, but in all other cases our selected DA outperforms the DA chosen by minimizing variance. The largest difference comes from Tiny-ImageNet MobileNetV2 when TrivialAugment outperforms CutMix by over 2.9%, and in CIFAR-100 VGG-13 when our chosen TrivialAugment outperforms AutoAugment ImageNet by 1.24%. In practice, the baseline approach selects a single DA for each dataset regardless of the teacher's model architecture. In contrast, our metric will select different DAs within a dataset, which we show to be particularly effective on CIFAR-100.

B. Improving SOTA KD-applied QAT

Results on CIFAR-10 and CIFAR-100. Tab. II shows that the proposed DA selection strategy significantly improves the existing QAT method (EWGS) and the KD-applied QAT (SQAKD) on CIFAR-10 and CIFAR-100, while also outperforming other existing KD-applied QAT benchmarks. The proposed DA selection strategy selects Trivial Augment on CIFAR-10 VGG-8 and CIFAR-100 VGG-13, RandAugment on CIFAR-10 ResNet-20, and AA ImageNet on CIFAR-100 ResNet-32 (see Sec. IV-A for details). The DA results in model improvements over EWGS and SQAKD in all cases except ResNet-32 W2A2. This is because the ResNet-32 model has relatively few parameters compared to VGG-13, and quantization to only 2 bits results in a model with insufficient capacity to generalize from the augmented data samples to the test dataset. We can see that the opposite occurs at the 4-bit quantization level, with the chosen DA yielding improvements of 1.5% over EWGS alone and 3.66% over SQAKD. This does appear to be part of a wider trend of the DA helping the W4A4 models more than the W2A2 models: the average improvement over SQAKD for the W2A2 quantized models is 0.72%, whereas the average for the W4A4 models is 2.78%.

Results on Tiny-ImageNet. Tab. III shows our performance on Tiny-ImageNet using our selected DA of Trivial Augment. In every case, our DA improves the results over using SQAKD or QAT alone. The benefit is extremely large in cases such as MobileNetV2 W4A4 where our DA improves top-1 accuracy of baseline PACT by 10.49%. Similarly, in ResNet-18 PACT W4A4, our DA improves over baseline SQAKD by 6.74%. Aside from the large benchmark improvements, the Tiny-ImageNet experiments are effective in showing that our selected DA works across a broad spectrum of QAT algorithms.

C. Improving SOTA KD Methods

Although we have focused on using standard KD thus far, Tab. IV demonstrates our method's efficacy in QAT+KD using other competitive KD algorithms. Our selected DA Trivial Augment successfully outperforms the baseline of standard data augmentations for all 8 evaluated methods. Each experiment is run for 200 epochs, with the loss weight hyperparameters set according to what is proposed in their original papers. Unless otherwise marked, every method includes loss from ground truth labels. Our DA selection has the largest impact on CRD and RLD, improving their accuracies by 3.2% and 1.94% respectively. Likewise, the NST, RKD, and SP algorithms are all improved by over 1% accuracy through the usage of the Trivial Augment. Interestingly, the smallest improvement is on KD without labels. This is likely because the teacher model is not explicitly trained under this DA, so the loss of labels may be more impactful than in the standard QAT+KD scenario. Nonetheless, the selected DA still improves the quantized model's performance, even if only by a small degree.

TABLE II: Top-1 test accuracy (%) on CIFAR-10 and CIFAR-100. "$W*A\times$" denotes that the weights and activations are quantized into $*$bit and $\times$bit, respectively. Full-precision models' accuracies are as follows: VGG-8 (92.91), ResNet-20 (94.00), VGG-13 (76.24), ResNet-32 (74.57). The results are subdivided into methods that use QAT only, and methods that use QAT and KD. EWGS [23], QKD [6], SPEQ [7], and SQAKD [5] results are cited from their original papers.

Dataset	CIFAR-10				CIFAR-100			
Model	VGG-8		ResNet-20		VGG-13		ResNet-32	
Bit-width	W2A2	W4A4	W2A2	W4A4	W2A2	W4A4	W2A2	W4A4
EWGS	90.84	90.95	91.41	92.40	73.31	73.41	**69.37**	70.50
Ours (EWGS)	**92.21**	**93.31**	**91.48**	**93.38**	**75.64**	**76.49**	67.84	**72.00**
QKD	-	-	90.5	93.1	-	-	66.4	73.3
SPEQ	-	-	91.4	-	-	-	69.1	-
SQAKD	91.55	91.31	91.80	92.59	74.65	74.67	**69.99**	71.65
Ours	**92.66**	**93.66**	**92.33**	**94.39**	**76.59**	**77.97**	69.27	**75.31**

TABLE III: Top-1 test accuracy (%) on Tiny ImageNet using MobileNetV2 and ResNet-18. QAT refers to quantization using QAT only, i.e. by PACT [21], LSQ [22], or DoReFa [18]. Both results are as reported in the SQAKD paper [5]. Our full-precision accuracies are: MobileNetV2 (58.64) and ResNet-18 (66.87)

	QAT	Ours (QAT)	SQAKD	Ours
MobileNetV2 PACT W3A3	47.77	**50.32**	52.73	**55.28**
MobileNetV2 PACT W4A4	50.33	**60.82**	57.14	**60.72**
MobileNetV2 DoReFa W8A8	56.26	**61.53**	58.13	**62.1**
ResNet-18 PACT W3A3	58.09	**58.4**	61.34	**65.45**
ResNet-18 LSQ W3A3	61.99	**65.56**	65.21	**67.34**
ResNet-18 PACT W4A4	61.06	**65.67**	61.47	**68.21**
ResNet-18 DoReFa W8A8	63.23	**66.85**	64.88	**68.28**

TABLE IV: Top-1 test accuracy (%) comparing our selected DA against the standard baseline DA across different methods of knowledge distillation using CIFAR-100 VGG-13 at W2A2. Full precision accuracy: 76.24%.

	Standard DA	Ours
AT [13]	75.49	**75.84**
CC [40]	73.51	**73.96**
CRD [10]	73.96	**77.16**
NST [14]	74.85	**76.33**
RKD [12]	74.05	**75.25**
RLD [16]	73.84	**75.78**
SP [41]	74.61	**76.29**
KD (no labels)	72.45	**72.66**

D. Empirical Performance of Quantized Models

Inference Latency. Fig. 1 presents measured speedups for 8-bit quantization across all evaluated architectures. We observe consistent improvements ranging from 2.5x for ResNet-20 on CIFAR-10 to 11.3x for ResNet-18 on TinyImageNet. The larger models exhibit greater relative speedups, with VGG-13 achieving 9.2x and ResNet-18 achieving 11.3x. MobileNetV2 shows more modest gains of 3.8x as its architecture is already optimized for efficiency. These results confirm that the effective QAT our DA selection method enables directly leads to substantial latency reductions.

Energy Consumption. Fig. 2 presents projected energy consumption based on the Horowitz model [43]. INT8 quantization reduces estimated energy by approximately 85% relative to FP32. This reduction is driven by both cheaper integer arithmetic, where an 8-bit multiply costs 0.2 pJ compared to 3.7 pJ for FP32, and reduced memory transfer costs. INT4 and INT2 show further reductions of approximately 93% and

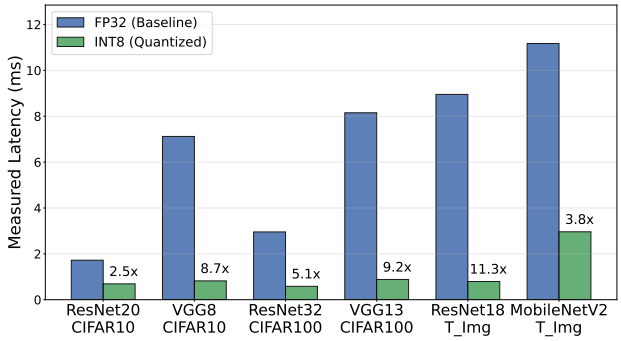
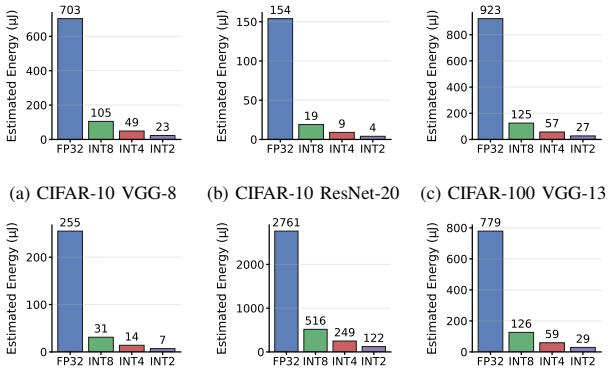


Fig. 1: Measured real world latency comparing full precision FP32 baselines against models quantized to W8A8. The relative speedup of the quantized model is listed above each bar.



(a) CIFAR-10 VGG-8 (b) CIFAR-10 ResNet-20 (c) CIFAR-100 VGG-13

(d) CIFAR-100 ResNet-32 (e) T_ImageNet MobileNet (f) T_ImageNet ResNet-18

Fig. 2: Estimated model energy consumption at different quantization levels across all surveyed model architectures. INT$*$ represents W$*$A$*$ quantization.

97%, respectively. These projections represent theoretical lower bounds; actual energy savings depend on hardware-accelerator support and memory-hierarchy effects. Nevertheless, the consistent pattern across architectures from the compact ResNet-20 to the larger VGG-13 demonstrates that our quantization translates to performance benefits that generalize across model families.

V. Ablation Study

Data Augmentation Selection. In Sec. IV-A, we showcase the performance of our DA selection metric using all 8 considered DAs. But, it is natural to ask whether our approach breaks down when the number of DAs being evaluated is relatively small. To this end, we evaluate our selection metric on all

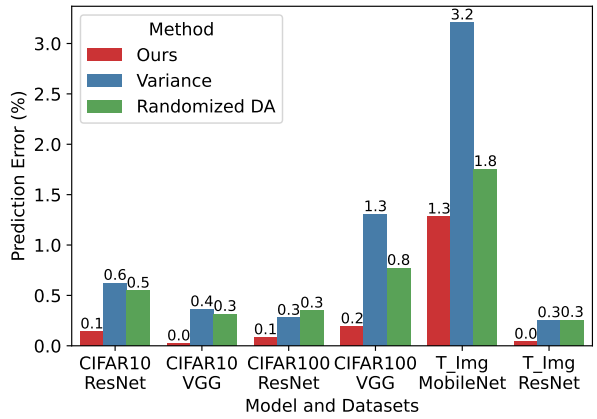
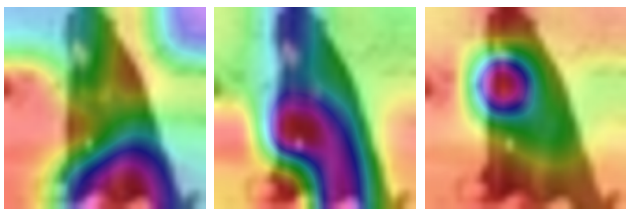


Fig. 3: Average prediction error for each method's selected DA against the optimal DA selection, averaged across all possible subsets the set of DAs. The models and datasets correspond to those included in Tab. I. T_Img is used as an abbreviation for Tiny-ImageNet.

non-trivial subsets of our 8 DAs that include the Standard DA. As before, we benchmark against the variance minimizing algorithm. Additionally, we compare against the randomized DA selection criteria, which reports the average performance of across all DAs except the Standard DA baseline. Fig. 3 displays the results of this DA selection ablation. In all cases, our DA selection metric outperforms both the variance and randomized DA baselines by a large margin. This both validates the conclusion of Sec. IV-A, and demonstrates that our method is not overly sensitive to the presence of specific DAs in its performance.

Activation Heatmap Visualization. Fig. 4 shows the LayerCam [44] heatmap superimposed on the image being classified. This visualization highlights areas of relative focus from red (low focus) to violet (high focus), using the activations in the convolutional filters of the network, with each point further scaled by its weight and gradient. The values across every convolutional filter are averaged, yielding a single color for each pixel. All three images are classified by models trained using the same datasets, but we can see that the model trained under our chosen DA (center) generates a heatmap that aligns much more closely with the teacher's (left). Although this is only a single example, this demonstrates that models trained to the same quantization level can exhibit vastly different behaviors depending on the DA they were trained under.



(a) Teacher (b) Trivial Augment (c) Standard DA

Fig. 4: The LayerCam [44] heatmaps for the teacher network (a), the quantized model with data augmentations (b), and the quantized model without data augmentations (c). The color indicates the area of focus of the network, on a scale of red (low focus) to violet (high focus).

Loss Terms. Tab. V shows our results with different hyper-parameter settings for loss terms trained under standard data augmentations. Our loss is a weighted sum between the cross entropy loss from labels (γ) and loss from KD (α)—in this case, we focus on traditional KD for simplicity. The case of $\gamma = 1, \alpha = 0$ corresponds to QAT only without KD. We find that this method is significantly worse than KD with QAT when $\alpha = 2$, matching the findings of previous work. In addition, we can see that in all experiments the model benefits from having labels available during training ($\gamma \neq 0$), similar to our findings in Sec. IV-C. Because we find the best configuration is $\gamma = 1, \alpha = 2$, we use this $1:2$ ratio for all our models trained using traditional KD.

TABLE V: Top-1 test accuracy (%) evaluated across hyperparameter settings for loss terms. α represents the weight applied to traditional KD loss and γ indicates weight attached to cross entropy loss from labels. These experiments were performed on CIFAR-100 using VGG-13 at W2A2 quantization.

	$\alpha = 0$	$\alpha = 1$	$\alpha = 2$
$\gamma = 0$	-	73.27	73.73
$\gamma = 1$	73.68	73.74	**74.67**

TABLE VI: Top-1 test accuracy (%) for quantized models under different initialization schemes. "Teacher" indicates the student is initialized to the teacher's parameters, "Random" indicates standard randomized model initialization. Both models are quantized to W2A2.

Initialization	Teacher	Random
VGG-13	**74.67**	66.79
ResNet-32	**70.62**	65.43

Network Initialization. Tab. VI shows the effects of network initialization for our QAT training. Because our teacher and student models are identical aside from the quantization, we can initialize the student's weights to the weights found in the teacher. We find that this initialization improves the resulting model performance upwards of 5% compared to random initialization, which is a unique benefit offered by the process of self-supervised distillation.

VI. Conclusion

In this paper, we propose a new metric for ranking data augmentations within the context of KD-assisted QAT. The metric is cheap to compute, and it can greatly reduce the search space for selecting a data augmentation. We establish the effectiveness of the metric by evaluating over several popular data augmentations, showing that our selected augmentation performs well, and that it consistently outperforms benchmarks set by existing works combining KD with QAT. We show the flexibility of the approach by evaluating across a range of model architectures, bit-widths, quantizers, and KD methods, demonstrating the utility of guided DA search for KD in QAT.

Acknowledgment

This project was supported in part by funding provided by the National Aeronautics and Space Administration (NASA), under award number 80NSSC25M7087, Michigan Space Grant Consortium (MSGC), as well as by the Oakland University Research Committee Award.

References

[1] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

[2] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3588–3597.

[3] B. Peng, X. Jin, J. Liu, S. Zhou, Y. Wu, Y. Liu, D. Li, and Z. Zhang, "Correlation congruence for knowledge distillation," 2019. [Online]. Available: https://arxiv.org/abs/1904.01802

[4] R. Krishnamoorthi, "Quantizing deep convolutional networks for efficient inference: A whitepaper," 2018. [Online]. Available: https://arxiv.org/abs/1806.08342

[5] K. Zhao and M. Zhao, "Self-supervised quantization-aware knowledge distillation," in *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, S. Dasgupta, S. Mandt, and Y. Li, Eds., vol. 238. PMLR, 02–04 May 2024, pp. 4375–4383. [Online]. Available: https://proceedings.mlr.press/v238/zhao24d.html

[6] J. Kim, Y. Bhalgat, J. Lee, C. Patel, and N. Kwak, "Qkd: Quantization-aware knowledge distillation," 2019. [Online]. Available: https://arxiv.org/abs/1911.12491

[7] Y. Boo, S. Shin, J. Choi, and W. Sung, "Stochastic precision ensemble: Self-knowledge distillation for quantized deep neural networks," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 8, pp. 6794–6802, May 2021. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/16839

[8] B. Zhuang, C. Shen, M. Tan, L. Liu, and I. Reid, "Towards effective low-bitwidth convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7920–7928.

[9] H. Wang, S. Lohit, M. N. Jones, and Y. Fu, "What makes a" good" data augmentation in knowledge distillation-a statistical perspective," *Advances in Neural Information Processing Systems*, vol. 35, pp. 13 456–13 469, 2022.

[10] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation distillation," 2022. [Online]. Available: https://arxiv.org/abs/1910.10699

[11] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015. [Online]. Available: https://arxiv.org/abs/1503.02531

[12] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[13] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," 2017. [Online]. Available: https://arxiv.org/abs/1612.03928

[14] Z. Huang and N. Wang, "Like what you like: Knowledge distill via neuron selectivity transfer," 2017. [Online]. Available: https://arxiv.org/abs/1707.01219

[15] B. Zhao, Q. Cui, R. Song, Y. Qiu, and J. Liang, "Decoupled knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 11 953–11 962.

[16] W. Sun, D. Chen, S. Lyu, G. Chen, C. Chen, and C. Wang, "Knowledge distillation with refined logits," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2025, pp. 1110–1119.

[17] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, and K. Keutzer, "A survey of quantization methods for efficient neural network inference," in *Low-power computer vision*. Chapman and Hall/CRC, 2022, pp. 291–326.

[18] S. Zhou, Y. Wu, Z. Ni, X. Zhou, H. Wen, and Y. Zou, "Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients," 2018. [Online]. Available: https://arxiv.org/abs/1606.06160

[19] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 or -1," 2016. [Online]. Available: https://arxiv.org/abs/1602.02830

[20] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "Xnor-net: Imagenet classification using binary convolutional neural networks," 2016. [Online]. Available: https://arxiv.org/abs/1603.05279

[21] J. Choi, Z. Wang, S. Venkataramani, P. I.-J. Chuang, V. Srinivasan, and K. Gopalakrishnan, "Pact: Parameterized clipping activation for quantized neural networks," 2018. [Online]. Available: https://arxiv.org/abs/1805.06085

[22] S. K. Esser, J. L. McKinstry, D. Bablani, R. Appuswamy, and D. S. Modha, "Learned step size quantization," 2020. [Online]. Available: https://arxiv.org/abs/1902.08153

[23] J. Lee, D. Kim, and B. Ham, "Network quantization with element-wise gradient scaling," 2021. [Online]. Available: https://arxiv.org/abs/2104.00903

[24] S. Jung, C. Son, S. Lee, J. Son, Y. Kwak, J.-J. Han, S. J. Hwang, and C. Choi, "Learning to quantize deep networks by optimizing quantization intervals with task loss," 2018. [Online]. Available: https://arxiv.org/abs/1808.05779

[25] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation policies from data," *arXiv preprint arXiv:1805.09501*, 2018.

[26] S. Lim, I. Kim, T. Kim, C. Kim, and S. Kim, "Fast autoaugment," 2019. [Online]. Available: https://arxiv.org/abs/1905.00397

[27] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 702–703.

[28] S. G. Müller and F. Hutter, "Trivialaugment: Tuning-free yet state-of-the-art data augmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 774–782.

[29] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "Augmix: A simple data processing method to improve robustness and uncertainty," *arXiv preprint arXiv:1912.02781*, 2019.

[30] H. Wang, S. Lohit, M. N. Jones, and Y. Fu, "What makes a "good" data augmentation in knowledge distillation - a statistical perspective," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 13 456–13 469. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/57b53238ff22bc0dc62de08f53eb5de2-Paper-Conference.pdf

[31] C. Pham, T. Hoang, and T.-T. Do, "Collaborative multi-teacher knowledge distillation for learning low bit-width deep neural networks," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2023, pp. 6435–6443.

[32] R. Krishnamoorthi, "Quantizing deep convolutional networks for efficient inference: A whitepaper," 2018. [Online]. Available: https://arxiv.org/abs/1806.08342

[33] J. H. Cho and B. Hariharan, "On the efficacy of knowledge distillation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4794–4802.

[34] L. Ye, S. M. Hamidi, R. Tan, and E.-H. Yang, "Bayes conditional distribution estimation for knowledge distillation based on conditional mutual information," *arXiv preprint arXiv:2401.08732*, 2024.

[35] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," 2019. [Online]. Available: https://arxiv.org/abs/1905.04899

[36] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," 2018. [Online]. Available: https://arxiv.org/abs/1710.09412

[37] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images.(2009)," 2009.

[38] Y. Le and X. Yang, "Tiny imagenet visual recognition challenge," *CS 231N*, vol. 7, no. 7, p. 3, 2015.

[39] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," 2015. [Online]. Available: https://arxiv.org/abs/1409.0575

[40] B. Peng, X. Jin, J. Liu, D. Li, Y. Wu, Y. Liu, S. Zhou, and Z. Zhang, "Correlation congruence for knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[41] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," 2019. [Online]. Available: https://arxiv.org/abs/1907.09682

[42] Y. Li, M. Shen, J. Ma, Y. Ren, M. Zhao, Q. Zhang, R. Gong, F. Yu, and J. Yan, "Mqbench: Towards reproducible and deployable model quantization benchmark," 2022. [Online]. Available: https://arxiv.org/abs/2111.03759

[43] M. Horowitz, "1.1 computing's energy problem (and what we can do about it)," in *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, 2014, pp. 10–14.

[44] P.-T. Jiang, C.-B. Zhang, Q. Hou, and Y. Wei, "Layercam: Exploring hierarchical class activation maps," *IEEE Transactions on Image Processing*, vol. PP, pp. 1–1, 06 2021.