CONTRASTIVE QUANT: QUANTIZATION MAKES STRONGER CONTRASTIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive learning, which learns visual representations by enforcing feature consistency under different augmented views, has emerged as one of the most effective unsupervised learning methods. In this work, we explore contrastive learning from a new perspective, inspired by the recent works showing that properly designed weight perturbations or quantization help the models learn a smoother loss landscape. Interestingly, we find that quantization, when properly engineered, can enhance the effectiveness of contrastive learning. To this end, we propose a novel contrastive learning framework, dubbed Contrastive Quant, to encourage the feature consistency under both (1) differently augmented inputs via various data transformations and (2) differently augmented weights/activations via various quantization levels. In Contrastive Quant, the feature consistency under injected noises via quantization can be viewed as augmentations on both the model weights and intermediate activations, serving as a complement to the input augmentations. Extensive experiments, built on top of two state-of-the-art contrastive learning methods SimCLR and BYOL, show that Contrastive Quant consistently improves the learned visual representation, especially under semi-supervised scenarios with limited labeled data. For example, our Contrastive Quant achieves a 8.69% and 10.27% higher accuracy on ResNet-18 and ResNet-34 with ImageNet, respectively, when fine-tuning with only 10% labeled data. We believe this work has opened up a new perspective for future contrastive learning innovations. All codes will be released upon acceptance.

1 INTRODUCTION

Contrastive learning has emerged as the state-of-the-art (SOTA) unsupervised representation learning from images. For example, Momentum Contrast (MoCo) (He et al., 2020) shows that unsupervised pre-training can surpass its ImageNet-supervised counterpart in multiple detection and segmentation tasks, while SimCLR further reduces the gap in linear classifier accuracy between unsupervised and supervised pre-training representations. As such, there has been a growing interest in further boosting its achievable performance and developing improved contrastive learning pipelines.

In this work, we explore contrastive learning from a new perspective, inspired by recent works showing that properly designed weight perturbations or quantization help the models learn a smoother loss landscape (Wu et al., 2020; Fu et al., 2021). For example, (Wu et al., 2020) adversarial perturbations on both inputs and weights help smooth the loss landscape of model weights and thus narrow the robust generalization gap and (Fu et al., 2021) shows that a properly designed precision schedule helps DNN converge to a better local optima, as a low precision helps the optimization space exploration in a similar way a high learning rate does. We are thus motivated to ask an intriguing question: "*Can quantization, which itself can boost the model efficiency, be leveraged to develop improved contrastive learning pipelines*"? If the answer is positive, it can not only lead to more accurate contrastive learning techniques on top of existing methods, but also open up a new understanding in the role of quantization on contrastive learning, potentially inspiring and motivating more contrastive learning innovations.

Interestingly, we find that quantization, when properly engineered, can enhance the effectiveness of contrastive learning. Specifically, we make the following contributions:

- We are the first to study the role of quantization in the context of contrastive learning pipelines, and show that quantization can be leveraged to enhance the performance of contrastive learning. We believe that this view can open up a new perspective for future contrastive learning innovations.
- We propose a novel contrastive learning framework, dubbed Contrastive Quant, to encourage the feature consistency under both (1) different augmented inputs via various data transformations and (2) different augmented weights/activations via various quantization levels. In particular, the feature consistency under injected noises via quantization in Contrastive Quant can be viewed as augmentations on both model weights and intermediate activations, serving as a complement to the input augmentations.
- Extensive experiments, built on top of two SOTA contrastive learning methods SimCLR and BYOL, show that our Contrastive Quant consistently improves the learned visual representation, especially with limited labeled data under semi-supervised scenarios. For example, our Contrastive Quant achieves a 8.69% and 10.27% higher accuracy on ResNet-18 and ResNet-34, respectively, on ImageNet when fine-tuning with 10% labeled data.

2 Related Works

DNN quantization. Quantization, which trims down the model complexity from the most fine-grained bit level, is one of the most promising DNN compression techniques. In particular, existing quantization methods represent model weights/activations/gradients using a lower floating-point precision (Wang et al., 2018; Sun et al., 2019) or fixed-point precision (Zhu et al., 2016; Li et al., 2016; Jacob et al., 2018; Mishra & Marr, 2017; Mishra et al., 2017; Park et al., 2017; Zhou et al., 2016). The resulting accuracy degradation after quantization can be minimized through (1) quantization-aware training (Jacob et al., 2018), which explicitly considers quantization noise in the training process, (2) learnable quantizers (Jung et al., 2019; Bhalgat et al., 2020; Esser et al., 2019; Park & Yoo, 2020), which jointly learn the quantization-related parameters and model parameters, and (3) mixed-precision quantization (Wang et al., 2019; Xu et al., 2018; Elthakeb et al., 2020; Zhou et al., 2017), which allocates different precisions to different layers. In addition to boosting efficiency, recent works (Fu et al., 2020; 2021) find that quantization can be properly utilized to boost training optimality. In particular, (Fu et al., 2021) shows that a low precision has a similar effect as a high learning rate, favoring the training space exploration and their proposed cyclic precision schedule helps DNNs converge to a better optima. This inspires and motivates us to explore the potential positive effects of quantization for representation learning.

Self-supervised learning and contrastive learning. Self-supervised learning, which leverages the input data themselves for supervision to benefit the downstream tasks, has achieved great progress. Early works (Van Oord et al., 2016; Oord et al., 2017; You et al., 2018; Radford et al., 2018; 2019; Kipf & Welling, 2016; Razavi et al., 2019; Devlin et al., 2018; Dinh et al., 2014; Mikolov et al., 2013a;b) adopt generative models to recover the original data distributions without making any assumptions for the downstream tasks to learn good representations. Later works shed light on the potential of discriminative models for representation learning. Many pioneering works along this direction aim at tailoring pretext tasks for different downstream tasks, and attempt to predict the missing information from intentionally manipulated training data, such as context prediction (Doersch et al., 2015), jigsaw puzzle (Noroozi & Favaro, 2016), colorization (Larsson et al., 2016; Zhang et al., 2016), rotation (Gidaris et al., 2018), and deep cluster (Caron et al., 2018; 2019). Such carefully designed pretext tasks are able to capture some common priors, which are also applicable to the downstream tasks. Recently, contrastive learning has gained increased popularity thanks to its excellent performance. The kernel spirit behind contrastive learning is to learn invariant features by maximizing the mutual information of the latent representations of differently augmented views of the images. In particular, instance discrimination (Wu et al., 2018) makes the first attempt to discriminate different instances via an Noise-Contrastive Estimation (NCE) loss and following works consider different strategies to construct the different views. For example, CMC (Tian et al., 2019) converts RGB images to the Lab color space and maximizes the mutual information between different color channel views, and SimCLR (Chen et al., 2020c) adopts different augmentations as different views and maximizes the consistency between different views. More variants of contrastive learning (He et al., 2020; Chen et al., 2020d; Grill et al., 2020; Chen & He, 2020) have been proposed to effectively generate negative pairs and improve the quality of representation learning. Readers are referred to (Liu et al., 2020; Jaiswal et al., 2021) for more details about unsupervised learning and contrastive learning.

In this work, we explore contrastive learning from a new perspective, i.e., the potential positive role of quantization in contrastive learning, and find that quantization, when properly engineered, can enhance the effectiveness of contrastive learning, potentially inspiring and motivating more contrastive learning innovations.

3 THE PROPOSED CONTRASTIVE QUANT FRAMEWORK

In this section, we introduce our Contrastive Quant framework, which for the first time explores quantization's positive effects on contrastive learning in addition to merely boosting the model efficiency. We start from the motivation of and inspiration leading to our framework in Sec. 3.1 and then introduce the key concept in Sec. 3.2. Next, we discuss potential designs to enhance the positive effect of quantization on contrastive learning and the implementation details for applying Contrastive Quant on top of existing contrastive learning frameworks in Sec. 3.3 and Sec. 3.4, respectively.

3.1 MOTIVATION

Spirits of contrastive learning. Self-supervised learning methods aim at learning representations with semantic priors that can generally benefit their downstream tasks. The core spirit behind contrastive learning (Tian et al., 2019; Wu et al., 2018; Chen et al., 2020c; He et al., 2020; Chen et al., 2020d; Grill et al., 2020; Chen & He, 2020), one of the most effective self-supervised learning methods, is to learn invariant features by maximizing the mutual information between the latent representations of differently augmented image views. Such a feature consistency is known to be beneficial for both the standard generalization (Zhang, 2019; Kayhan & Gemert, 2020) and robust generalization (Kim et al., 2020; Chen et al., 2020b; Jiang et al., 2020), while new perspectives for enforcing such feature consistency in addition to data augmentations are still under-explored.

Inspirations from recent works. A recent work (Wu et al., 2020) implies another view of encouraging feature consistency in the context of adversarial training. In particular, they show that training with properly generated perturbations onto the weights can serve as a complement for adversarial perturbations onto the inputs, which helps smooth the loss landscape and narrow the robust generation gap, i.e., improve the feature consistency under adversarial attacks. In parallel, (Fu et al., 2021) shows that a low precision has a similar effect as a high learning rate, favoring the training space exploration, and proposes a properly designed precision schedule to help DNN converge to a better local optima with more smoothed loss landscape. Considering quantization can naturally serve as perturbations onto both model weights and intermediate feature maps, these prior works inspire and motivate us to leverage quantization to learn better representations by encouraging feature consistency under differently augmented weights/activations via various quantization levels, as a complement of enforcing consistency via data augmentations.

3.2 THE KEY CONCEPT

Contrastive learning, as a kind of instance discrimination method, encourages feature consistency via maximizing the mutual information, approximated by minimizing an NCE loss (Wu et al., 2018), between the extracted features from different perspectives of the same instance which can be formulated as:

$$\max\ I(f, f^+) \approx \min\ NCE(f, f^+) = \min\ \mathbb{E}\,[-log\frac{exp(f \cdot f^+/\tau)}{\sum_{i=0}^{K} exp(f \cdot f^-/\tau) + exp(f \cdot f^+/\tau)}] \quad (1)$$

where f and f^- are the extracted features of the same instance under different perspectives, named positive pairs, of the encoder F, $I(f, f^+)$ denotes their mutual information, K is the number of negative samples, and τ is a temperature parameter that controls the concentration level of the distribution.

In previous data augmentation based contrastive learning methods, the positive pairs f and f^+ are generated by different augmentation combinations:

$$f = F(Aug_1(x), \theta)\ ,\ \ f^+ = F(Aug_2(x), \theta) \quad (2)$$

where x is the given instance, θ is the model weight, and Aug_1 and Aug_1 denote two different augmentations. To further enhance the feature consistency from a new perspective, we propose the Contrastive Quant framework to augment both the model weights and intermediate activations in

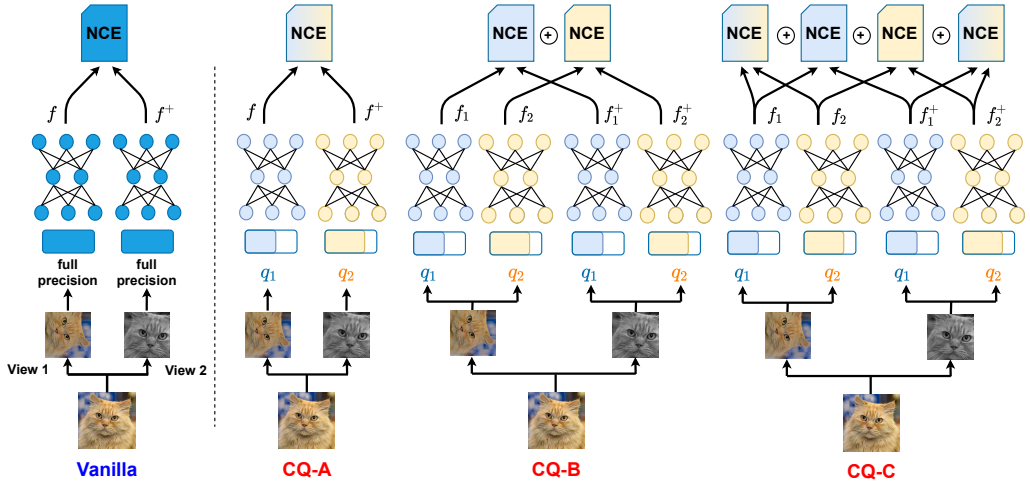

Figure 1: An overview of the proposed Contrastive Quant design pipelines.

addition to the input augmentations via injecting quantization noise of different levels into the weights and activations, as formulated below:

$$f = F_{q_1}(x, \theta_{q_1}) \ , \ f^+ = F_{q_2}(x, \theta_{q_2}) \tag{3}$$

where f and f^+ are generated by the encoder quantized to different precisions q_1 and q_2, which can be randomly selected from a precision set during training. The detailed quantization scheme we adopt is discussed in Sec. 3.4.

Nevertheless, we empirically find that merely enhancing the feature consistency via quantization augmented weights/activations without the vanilla input augmentations will lead to inferior contrasting learning, indicating that the structured priors learned from the consistency between different augmented views are essential to the downstream tasks and it's necessary to apply our Contrastive Quant framework on top of existing input augmentation based contrasting learning methods. A natural question is thus "how to effectively combine augmented inputs and augmented weights/activations via quantization to boost the performance of contrastive learning"? To answer this question, we explore the potential designs in Sec. 3.3 towards an effective Contrastive Quant framework.

3.3 The Design Pipeline

As shown in Fig. 1, we propose three candidate designs, denoted as CQ-A, CQ-B, and CQ-C, respectively, to exploit the potential of applying augmented weights/activations via quantization towards on top of the commonly used augmented inputs towards better contrastive learning pipelines.

Analysis of CQ-A. CQ-A (see the second column of Fig. 1) is one of the most intuitive designs for applying augmented weights/activations on top of existing input augmentation based contrastive learning methods, which can be formulated as:

$$f = F_{q_1}(Aug_1(x), \theta_{q_1}) \ , \ f^+ = F_{q_2}(Aug_2(x), \theta_{q_2}) \ , \ Loss = NCE(f, f^+) \tag{4}$$

where $Loss$ is the final objective to be minimized. Eq. 4 shows that CQ-A views the quantization precision as an additional augmentation parameter similar to the rotation degree or noise level in data augmentation operators (Chen et al., 2020c), which are randomly selected during inference with different views for the same instance. In this way, the executions of input augmentations and weight/activation augmentation are combined in a sequential manner, which can potentially strengthen the overall augmentation magnitude. As validated in Sec. 4, such strengthened augmentations will aggressively benefit large-scale datasets like ImageNet while may not be very helpful on small-scale datasets as strong augmentations may distort the images' structures (Wang & Qi, 2021), which is more likely to happen on small-scale datasets.

Analysis of CQ-B. CQ-B (see the third column of Fig. 1) is a variant with more mild augmentations. Instead of explicitly constraining the feature consistency under differently augmented

weights/activations via various quantization levels, CQ-B only enforces the feature consistency under differently augmented inputs with the same randomly selected precision, which can be formulated as:

$$f_1 = F_{q_1}(Aug_1(x), \theta_{q_1}), \ f_2 = F_{q_2}(Aug_1(x), \theta_{q_2}) \tag{5}$$

$$f_1^+ = F_{q_1}(Aug_2(x), \theta_{q_1}), \ f_2^+ = F_{q_2}(Aug_2(x), \theta_{q_2}) \tag{6}$$

$$Loss = NCE(f_1, f_1^+) + NCE(f_2, f_2^+) \tag{7}$$

The final objective is averaged over two randomly selected precisions, which implicitly encourages the feature consistency under different quantization levels. As such, this design can potentially mitigate the potential risk of distorting the images' structures with too strong augmentations, especially on small-scale datasets, while the newly introduced priors by CQ-B over input augmentation based methods are not rich enough, which can limit its achievable performance improvement.

Analysis of CQ-C. To combine the advantages of both CQ-A and CQ-B, CQ-C (see the rightmost column of Fig. 1) explicitly encourages (1) the feature consistency under differently augmented views with the same quantization level, and (2) the feature consistency under differently augmented weights/activations via various quantization levels within the same view, which can be formulated as:

$$Loss = NCE(f_1, f_1^+) + NCE(f_2, f_2^+) + NCE(f_1, f_2) + NCE(f_1^+, f_2^+) \tag{8}$$

where f_1, f_2, f_1^+, and f_2^+ follow the definition in Eq. 5 and 6. Different from CQ-A which sequentially augments the inputs and weights/activations, CQ-C decouples them and enforce the feature consistency from the two perspectives separately, which can potentially mitigate the risk of introducing too strong augmentations while introducing sufficient new priors on top of existing contrastive learning methods. As validated in Sec. 4, CQ-C can consistently improve the performance on both small-scale and large-scale datasets, especially outperforming CQ-A on small-scale ones.

3.4 More implementation details

The adopted quantization scheme. We adopt the commonly adopted linear quantizer (Jacob et al., 2018) to quantize both weights and activations in our Contrastive Quant, i.e.,

$$A_q = S_a \lfloor \frac{A}{S_a} \rfloor \ , \quad where \ S_a = \frac{A_{range}}{2^q - 1} \tag{9}$$

where A here can denote the model weights or activations, q is the quantization bit-width, and A_{range} is the dynamic range of A, i.e., the difference between the maximum and minimum of A values. Since the encoder will be quantized to different values during training, learnable quantizers with trainable quantization parameters are found to be unstable here, so we directly adopt the linear quantizer.

Applying on top of SimCLR. Adapting our Contrastive Quant to the SOTA SimCLR framework (Chen et al., 2020c) is simple and direct via (1) modifying the NCE loss in Eq. 1 to the NT-Xent one (Chen et al., 2020c), and (2) adding a projection head after the encoder to learn a better representation.

Applying on top of BYOL. BYOL (Grill et al., 2020) relies on two networks, i.e., the online and target networks, to learn from each other based on the feature consistency under different views, where the target network is updated via moving average of the online network instead of the gradients, and this training process does not involve any negative pair. We adapt our Contrastive Quant in a natural manner in that (1) we modify the NCE loss in Eq. 1 to the Mean Square Error (MSE) loss adopted by (Grill et al., 2020); (2) we add a projection head and prediction head after the encoder following (Grill et al., 2020); and (3) we stop the gradient propagation along the target network and apply both views of the same instance into the online network/target network alternatively to improve the data reusability following (Grill et al., 2020).

4 Experiment Results

In this section, we first introduce our experiment setup, benchmarking experiment results over SOTA contrastive learning methods, and then ablation studies of Contrastive Quant for better understanding its effectiveness and design pipelines.

Table 1: Benchmark Constrastive Quant against SimCLR on top of ResNet-18/34 on ImageNet. Here we adopt the fine-tuning settings on 10%/1% labeled data with the two precision sets.

Network	Method	Precision Set	Fine-tune Acc. (FP)		Fine-tune Acc. (4-bit)	
			10% labels	1% labels	10% labels	1% labels
ResNet-18	SimCLR	-	42.44	19.18	39.12	17.24
	CQ-A	6-16	**51.39**	28.87	**48.80**	**27.13**
		8-16	51.13	**28.97**	48.63	26.66
	CQ-C	6-16	44.97	20.83	42.01	18.63
		8-16	45.10	20.98	41.90	18.72
ResNet-34	SimCLR	-	47.53	23.43	44.65	21.69
	CQ-A	6-16	**55.76**	33.37	53.32	31.30
		8-16	55.72	**33.70**	**53.33**	**31.64**
	CQ-C	6-16	50.45	26.32	47.65	24.53
		8-16	50.22	26.21	47.70	24.74

4.1 Experiment setup

Networks, datasets, and evaluation settings. We consider six networks on two datasets, i.e., ResNet-18/34/74/110/156/MobileNetV2 on CIFAR-100 (Krizhevsky et al., 2009) and ResNet-18/34 on ImageNet (Deng et al., 2009), featuring diverse DNN models and data statistics for a solid evaluation. We also transfer the pretrained models on ImageNet to the downstream detection task Pascal VOC (Everingham et al., 2010; 2015). In this work, we consider both the fine-tuning, linear evaluation, and transfer learning settings. As we quantize the model to different precisions during the contrastive learning processes, we mainly consider the fine-tuning setting under a fixed precision, i.e.,, full precision (denoted as FP) or 4-bit, with a limited amount of labeled data (10% or 1%) to stabilize the weight/activation distribution under the precision choice. Detailed training and evaluation settings are provided in Appendix. A.

Precision sets. As our Contrastive Quant framework randomly selects two precision q_1 and q_2 from a pre-defined precision set in each training iteration, the precision set may influence its training optimality. We adopt *4-16* (every precision between 4-bit and 16-bit), *6-16*, and *8-16* as the potential precision sets for an ablation study in Appendix. B.

4.2 Benchmark Contrastive Quant on ImageNet

We first apply our Contrastive Quant framework on top of SimCLR to train ResNet-18/34 on ImageNet and benchmark with the vanilla SimCLR in both fine-tuning and linear evaluation settings under both FP and 4-bit precisions. In particular, we validate the effectiveness of all the three designs in Fig. 1 and adopt two precision sets *6-16* and *8-16* considering 4-bit may significantly degrade the final accuracy on large-scale datasets like ImageNet.

Fine-tuning results. As shown in Tab. 1, we can see that (1) both CQ-A and CQ-C achieve a consistently improvement over the vanilla SimCLR on both ResNet-18/34 and the two precision sets, indicating enforcing feature consistency under differently augmented weights/activations via various quantization levels indeed benefits the downstream tasks, which could provide a new prior; (2) CQ-C achieves a 1.39%∼2.89% and 2.69%∼3.05% higher accuracy over SimCLR on ResNet-18 and ResNet-34, respectively, while CQ-A achieves an even surprising at of 8.69%∼9.89% and 8.19%∼10.27% on ResNet-18 and ResNet-34, respectively, indicating the sequentially applied augmentations onto the inputs and weights/activations, which together lead to a stronger augmentation, greatly benefit the training process on large-scale datasets; and (3) we find that the training process of CQ-B can easily fail, which suffers from severe gradient explosion before learning a good representation, indicating the importance of adopting the feature consistency loss of differently augmented weights/actions under the same view as CQ-C in Sec. 3.3, which is the only difference between the two designs.

Linear evaluation results. As observed from Tab. 2, we can see that CQ-C and CQ-A achieve a 2.59%/1.18% and 15.60%/12.92% higher accuracy on ResNet-18/34, respectively. Furthermore, CQ-A still achieves the most aggressive improvements over SimCLR, aligning with our assumption about CQ-A's stronger augmentation effects which benefits ImageNet training.

Table 2: Comparing CQ-A and CQ-C with SimCLR under the linear evaluation setting on ImageNet.

Network	SimCLR	CQ-C	**CQ-A**
ResNet-18	29.31	31.90	**44.91**
ResNet-34	34.96	36.14	**47.88**

Table 4: Benchmark against SimCLR on six network with CIFAR-100 with the fine-tuning setting.

Network	Method	Fine-tune Acc. **(FP)**		Fine-tune Acc. **(4-bit)**	
		10% labels	**1%** labels	**10%** labels	**1%** labels
ResNet-18	SimCLR	61.51	42.51	59.78	40.73
	CQ-C	**61.75**	**43.80**	**60.12**	**42.59**
ResNet-34	SimCLR	63.05	45.11	61.44	43.63
	CQ-C	**63.58**	**48.05**	**61.47**	**45.75**
ResNet-74	SimCLR	51.93	30.40	50.37	28.56
	CQ-C	**52.52**	**31.39**	**51.12**	**29.70**
ResNet-110	SimCLR	52.78	31.16	51.69	30.11
	CQ-C	**54.47**	**33.17**	**52.28**	**32.66**
ResNet-152	SimCLR	53.57	32.93	52.14	31.06
	CQ-C	**55.44**	**34.98**	**53.04**	**33.54**
MobileNetV2	SimCLR	49.73	24.18	46.47	18.98
	CQ-C	**51.59**	**26.12**	**49.82**	**20.82**

Table 5: Benchmark against SimCLR on six DNNs & CIFAR-100 under the linear evaluation setting.

Method	ResNet-18	ResNet-34	ResNet-74	ResNet-110	ResNet-152	MobileNetV2
SimCLR	**64.91**	65.92	52.96	53.53	53.97	52.53
CQ-C	64.78	**66.54**	**54.06**	**54.76**	**55.12**	**53.97**

Transfer to downstream detection tasks. We further transfer the pretrained ResNet-18/34 trained with different methods on ImageNet to a downstream detection task on top of YOLOv4 (Bochkovskiy et al., 2020) following (Chen et al., 2020a). In particular, we follow (He et al., 2020; Chen et al., 2020d;a) to train the models on the combined training/validation set of Pascal VOC 2007 (Everingham et al., 2010) and Pascal VOC 2012 (Everingham et al., 2015), and evaluate on the test set of Pascal VOC 2007. As observed from Tab. 3, we can see that CQ-A and CQ-C consistently outperform the vanilla SimCLR after being transferred to the downstream detection task in terms of all the three metrics. Specifically, CQ-A achieves a 11.30%/3.19% higher AP compared with the vanilla SimCLR on top of ResNet-18/34, respectively.

Table 3: Comparing CQ-A and CQ-C with vanilla SimCLR via transferring the ImageNet pretrained models to the downstream detection task.

Network	Method	AP	AP50	AP75
ResNet-18	Vanilla SimCLR	25.09	49.2	22.74
	CQ-C	32.94	63.96	29.28
	CQ-A	**36.39**	**69.08**	**32.64**
ResNet-34	Vanilla SimCLR	35.58	67.51	31.88
	CQ-C	36.54	68.77	34.17
	CQ-A	**38.77**	**72.13**	**35.85**

Insights. (Wang & Qi, 2021) finds that stronger augmentations can distort the images' structures (Wang & Qi, 2021), thus can be harmful to learn a good representation, while they mainly discuss within the domain of input augmentations. Based on the success of CQ-A in Tab. 1, our Contrastive Quant can potentially serve as a new direction to explore stronger augmentations which consistently benefit the downstream tasks. A future direction is to explore other kinds of perturbations on weights/activations in addition to our Contrastive Quant to build more effective augmentations.

4.3 Benchmark Contrastive Quant on CIFAR-100

We then benchmark Contrastive Quant on top of SimCLR (Chen et al., 2020c)/BYOL (Grill et al., 2020) with the vanilla SimCLR/BYOL. As shown in Sec. 4.4, CQ-C outperforms CQ-A and CQ-B on small-scale datasets like CIFAR-100, which is consistent with the analysis in Sec. 3.3. Therefore, in this subsection we adopt CQ-C with a precison set of *6-16* and the results on more precision sets are discussed in Appendix. B.

Benchmark against SimCLR with fine-tuning settings. As shown in Tab. 4, we can see that (1) our CQ-C consistently outperforms the vanilla SimCLR on all the six networks, e.g., an accuracy improvement of 0.24%~3.35% and 0.99%~2.94% when fine-tuning with 10% and 1% data with FP,

Table 6: Benchmark against BYOL on three network with CIFAR-100 under the fine-tuning setting.

Network	Method	Precision Set	Fine-tune Acc. (**FP**) **10%** labels	Fine-tune Acc. (**FP**) **1%** labels	Fine-tune Acc. (**4-bit**) **10%** labels	Fine-tune Acc. (**4-bit**) **1%** labels
ResNet-18	BYOL	-	55.26	34.22	53.44	32.93
	CQ-C	6-16	58.84	39.21	56.74	37.54
ResNet-34	BYOL	-	65.83	50.95	64.00	49.37
	CQ-C	6-16	66.77	51.91	65.21	50.55
MobileNetV2	BYOL	-	49.85	23.32	44.65	19.58
	CQ-C	6-16	54.59	31.96	50.97	26.60

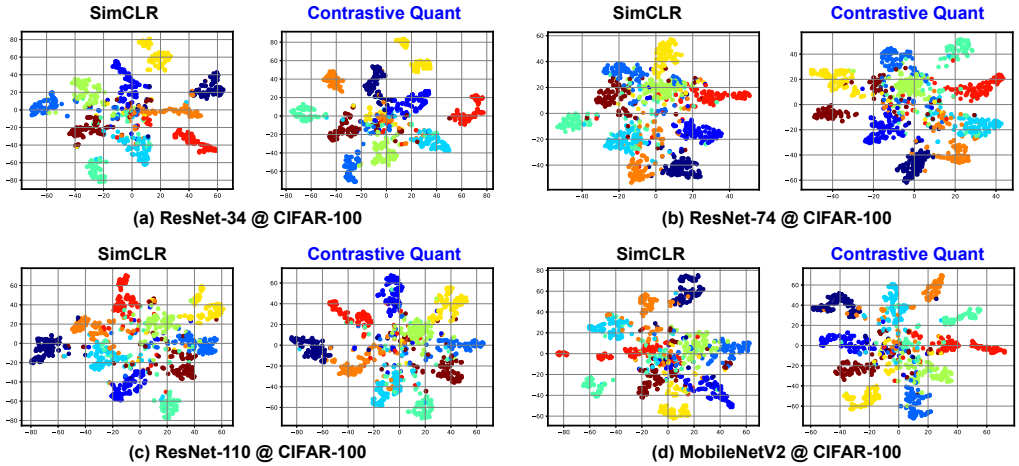

Figure 2: Visualizing the learned representations of Contrastive Quant and SimCLR using t-SNE (Van der Maaten & Hinton, 2008).

respectively; and (2) our CQ-C achieves more notable improvements over SimCLR on top of larger models and less labeled data, indicating its scalability and practicality in the real-world applications.

Benchmark against SimCLR with linear evaluation settings. Tab. 5 shows that our CQ-C still achieves a better accuracy (except ResNet-18) than the vanilla SimCLR, especially on larger networks.

Benchmark against BYOL with fine-tuning settings. A consistent improvement can be observed when benchmarking against BYOL in Tab. 6, i.e., CQ-C achieves a 0.94%~6.32% and 0.96%~8.64% higher accuracy when fine-tuning with 10% and 1% data with FP, respectively.

Visualizing the learned representations. We adopt t-SNE (Van der Maaten & Hinton, 2008) to visualize the learned representations (before the projection head) of the models trained by Contrastive Quant and SimCLR in Tab. 4 and Fig. 2. We can see that the representations learned by Contrastive Quant show a better linear separability, especially under larger models.

4.4 Ablation Studies: Benchmark Different Designs

As shown in Tab. 7, we can observe that (1) CQ-C achieves an overall better performances than the other two variants, especially when be fine-tuned with less labeled data, and (2) CQ-A generally achieves marginally better or comparable results over the vanilla SimCLR, which is different from its superior performance on ImageNet in Tab. 1, which aligns with our analysis in Sec. 3.3 that the too strong augmentations may distort the images' structures (Wang & Qi, 2021) small-scale datasets.

4.5 Ablation Studies: Augment with Different Quantization Levels Only

Setup. To justify whether different quantization levels in our Contrastive Quant play a similar role as data augmentations, we design a new variant of Contrastive Quant named CQ-quant, where each input is only augmented by different quantization levels without being augmented by different

Table 7: Ablation studies of Constrastive Quant's variants with a precision set of *6-16* on CIFAR-100.

Network	**Method**	Fine-tune Acc. **(FP)**		Fine-tune Acc. **(4-bit)**	
		10% labels	**1%** labels	**10%** labels	**1%** labels
ResNet-34	SimCLR	63.05	45.11	61.44	43.63
	CQ-A	**63.63**	45.60	**61.77**	43.56
	CQ-B	63.57	45.26	61.76	43.60
	CQ-C	63.58	**48.05**	61.47	**45.75**
ResNet-74	SimCLR	51.93	30.40	50.37	28.56
	CQ-A	51.89	29.95	**51.45**	28.99
	CQ-B	52.36	30.48	51.20	29.28
	CQ-C	**52.52**	**31.39**	51.12	**29.70**
MobileNetV2	SimCLR	49.73	24.18	46.47	18.98
	CQ-A	49.93	24.57	46.01	19.38
	CQ-B	**51.78**	25.21	47.81	20.81
	CQ-C	51.59	**26.12**	**49.82**	**20.82**

Table 8: Evaluating CQ-Quant augmented by different quantization levels only on CIFAR-100.

Network	**Precision Set**	Fine-tune Acc. (**FP**)		**Linear evaluation**
		1% labels	**10%** labels	
ResNet-74	6-16	**7.64**	**29.14**	**15.79**
	8-16	4.64	21.37	10.98
	No SSL Training	2.90	20.76	3.69
ResNet-110	6-16	**7.43**	**27.69**	**14.10**
	8-16	6.41	21.58	11.83
	No SSL Training	2.21	20.56	3.15

data augmentation methods. To be more specific, the loss function is modified from Eq. 8 to be: $Loss = NCE(f_1, f_2)$. We benchmark CQ-Quant with the baseline without SSL training, i.e., training from scratch during evaluation, on top of ResNet-74 and ResNet-110.

Observations. As shown in Tab. 8, we can observe that (1) CQ-Quant with different precision sets can consistently outperform the baseline without SSL training, indicating that the different quantization levels can indeed create a contrastive task with a similar effect as different data augmentions on the inputs; (2) CQ-Quant with more diverse precision settings can achieve a better fine-tuning/linear-evaluation accuracy, which aligns with our intuition that the diversity of augmentations contributes to the success of contrastive learning; and (3) data augmentations are still necessary towards decent contrastive learning performances.

5 Conclusion

In this work, we for the first time explore quantization's positive effects on boosting contrastive learning's accuracy in addition to merely boosting the model efficiency, and propose a novel contrastive learning framework, dubbed Contrastive Quant, to enhance the feature consistency under both (1) differently augmented inputs via various data transformations and (2) differently augmented weights/activations via various quantization levels. Extensive experiments on top of two state-of-the-art contrastive learning methods, SimCLR and BYOL, show that Contrastive Quant consistently improves the learned visual representation, especially with limited labeled data under semi-supervised scenarios. Our Contrastive Quant can not only lead to more accurate contrastive learning techniques on top of existing methods, but also open up a new understanding in the role of quantization on contrastive learning, potentially inspiring and motivating more contrastive learning innovations.

REFERENCES

Yash Bhalgat, Jinwon Lee, Markus Nagel, Tijmen Blankevoort, and Nojun Kwak. Lsq+: Improving low-bit quantization through learnable offsets and better initialization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 696–697, 2020.

Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 132–149, 2018.

Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin. Unsupervised pre-training of image features on non-curated data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2959–2968, 2019.

Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Michael Carbin, and Zhangyang Wang. The lottery tickets hypothesis for supervised and self-supervised pre-training in computer vision models. *arXiv preprint arXiv:2012.06908*, 2020a.

Tianlong Chen, Sijia Liu, Shiyu Chang, Yu Cheng, Lisa Amini, and Zhangyang Wang. Adversarial robustness: From self-supervised pre-training to fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 699–708, 2020b.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020c.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *arXiv preprint arXiv:2011.10566*, 2020.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020d.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pp. 1422–1430, 2015.

Ahmed Taha Elthakeb, Prannoy Pilligundla, Fatemeh Mireshghallah, Amir Yazdanbakhsh, and Hadi Esmaeilzadeh. Releq: A reinforcement learning approach for automatic deep quantization of neural networks. *IEEE Micro*, 2020.

Steven K Esser, Jeffrey L McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S Modha. Learned step size quantization. *arXiv preprint arXiv:1902.08153*, 2019.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.

Yonggan Fu, Haoran You, Yang Zhao, Yue Wang, Chaojian Li, Kailash Gopalakrishnan, Zhangyang Wang, and Yingyan Lin. Fractrain: Fractionally squeezing bit savings both temporally and spatially for efficient dnn training. *arXiv preprint arXiv:2012.13113*, 2020.

Yonggan Fu, Han Guo, Meng Li, Xin Yang, Yining Ding, Vikas Chandra, and Yingyan Lin. Cpt: Efficient deep neural network training via cyclic precision. *arXiv preprint arXiv:2101.09868*, 2021.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2704–2713, 2018.

Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2021.

Ziyu Jiang, Tianlong Chen, Ting Chen, and Zhangyang Wang. Robust pre-training by adversarial contrastive learning. *arXiv preprint arXiv:2010.13337*, 2020.

Sangil Jung, Changyong Son, Seohyung Lee, Jinwoo Son, Jae-Joon Han, Youngjun Kwak, Sung Ju Hwang, and Changkyu Choi. Learning to quantize deep networks by optimizing quantization intervals with task loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4350–4359, 2019.

Osman Semih Kayhan and Jan C van Gemert. On translation invariance in cnns: Convolutional layers can exploit absolute spatial location. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14274–14285, 2020.

Minseon Kim, Jihoon Tack, and Sung Ju Hwang. Adversarial self-supervised contrastive learning. *arXiv preprint arXiv:2006.07589*, 2020.

Thomas N Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.

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

Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *European conference on computer vision*, pp. 577–593. Springer, 2016.

Fengfu Li, Bo Zhang, and Bin Liu. Ternary weight networks. *arXiv preprint arXiv:1605.04711*, 2016.

Xiao Liu, Fanjin Zhang, Zhenyu Hou, Zhaoyu Wang, Li Mian, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *arXiv preprint arXiv:2006.08218*, 1(2), 2020.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*, 2013b.

Asit Mishra and Debbie Marr. Apprentice: Using knowledge distillation techniques to improve low-precision network accuracy. *arXiv preprint arXiv:1711.05852*, 2017.

Asit Mishra, Eriko Nurvitadhi, Jeffrey J Cook, and Debbie Marr. Wrpn: wide reduced-precision networks. *arXiv preprint arXiv:1709.01134*, 2017.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pp. 69–84. Springer, 2016.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*, 2017.

Eunhyeok Park and Sungjoo Yoo. Profit: A novel training method for sub-4-bit mobilenet models. *arXiv preprint arXiv:2008.04693*, 2020.

Eunhyeok Park, Junwhan Ahn, and Sungjoo Yoo. Weighted-entropy-based quantization for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5456–5464, 2017.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *arXiv preprint arXiv:1906.00446*, 2019.

Xiao Sun, Jungwook Choi, Chia-Yu Chen, Naigang Wang, Swagath Venkataramani, Vijayalakshmi Viji Srinivasan, Xiaodong Cui, Wei Zhang, and Kailash Gopalakrishnan. Hybrid 8-bit floating point (hfp8) training and inference for deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 4901–4910, 2019.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

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

Aaron Van Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International Conference on Machine Learning*, pp. 1747–1756. PMLR, 2016.

Kuan Wang, Zhijian Liu, Yujun Lin, Ji Lin, and Song Han. Haq: Hardware-aware automated quantization with mixed precision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8612–8620, 2019.

Naigang Wang, Jungwook Choi, Daniel Brand, Chia-Yu Chen, and Kailash Gopalakrishnan. Training deep neural networks with 8-bit floating point numbers. In *Advances in neural information processing systems*, pp. 7675–7684, 2018.

Xiao Wang and Guo-Jun Qi. Contrastive learning with stronger augmentations. *arXiv preprint arXiv:2104.07713*, 2021.

Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 33, 2020.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3733–3742, 2018.

Yuhui Xu, Shuai Zhang, Yingyong Qi, Jiaxian Guo, Weiyao Lin, and Hongkai Xiong. Dnq: Dynamic network quantization. *arXiv preprint arXiv:1812.02375*, 2018.

Jiaxuan You, Bowen Liu, Rex Ying, Vijay Pande, and Jure Leskovec. Graph convolutional policy network for goal-directed molecular graph generation. *arXiv preprint arXiv:1806.02473*, 2018.

Richard Zhang. Making convolutional networks shift-invariant again. In *International Conference on Machine Learning*, pp. 7324–7334. PMLR, 2019.

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pp. 649–666. Springer, 2016.

Shuchang Zhou, Yuxin Wu, Zekun Ni, Xinyu Zhou, He Wen, and Yuheng Zou. Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. *arXiv preprint arXiv:1606.06160*, 2016.

Yiren Zhou, Seyed-Mohsen Moosavi-Dezfooli, Ngai-Man Cheung, and Pascal Frossard. Adaptive quantization for deep neural network. *arXiv preprint arXiv:1712.01048*, 2017.

Chenzhuo Zhu, Song Han, Huizi Mao, and William J Dally. Trained ternary quantization. *arXiv preprint arXiv:1612.01064*, 2016.

A TRAINING AND EVALUATION SETTINGS

Training & fine-tuning & linear evaluation on ImageNet. We follow the training settings in (Chen et al., 2020c) to adopt a LARS optimizer and a cosine learning rate decay for training on ImageNet. In particular, we adopt a batch size of 512 with an initial learning rate of 0.6 to train the models for 100 epochs. For fine-tuning the pretrained models, we follow (Chen et al., 2020c) to adopt an SGD optimizer with Nesterov momentum and a momentum parameter of 0.9 without any weight decay. Specifically, we fine-tune the pretrained models using a batch size of 512 and a cosine learning rate decay with an initial learning rate of 0.1 for totally 30 epochs on 10% labeled data or 60 epochs on 1% labeled data. For the linear evaluation, we follow (Chen et al., 2020c) to train a linear classifier on top of the encoded features for 90 epochs using a Nesterov momentum optimizer with a momentum of 0.9, a learning rate of 0.2, and a batch size of 512 without weight decay.

Training & fine-tuning & linear evaluation on CIFAR-100. We follow all the settings in (Chen et al., 2020c) and (Jiang et al., 2020), including the optimizer settings, augmentation choices, and the projection head design. We train the models with a batch size of 512 for 1000 epochs. For both fine-tuning and linear evaluation settings, we adopt an SGD optimizer with a momentum of 0.9 and a cosine learning rate decay with an initial learning rate of 0.1 to fine-tune the models for 50 epochs.

Training on Pascal VOC. We follow (Chen et al., 2020a) to train each model for 50 epochs using an SGD optimizer with a momentum of 0.9 and a weight decay of 0.0005, and a cosine learning rate decay with an initial learning rate of 0.0001 and a batch size of 8. The evaluation metrics AP, AP50, and AP75 (Chen et al., 2020d) are all reported.

B ABLATION STUDIES: THE INFLUENCE OF PRECISION SETS

We also conduct an ablation study for evaluating Contrastive Quant's performance dependence on the adopted precision sets in Tab. 9 and find that (1) our Contrastive Quant generally achieves a decent performance under all considered precision sets and there's no gold precision set choice, and (2) generally training with low precision sets (*4-16*) will benefit Contrastive Quant's performance under the fine-tuning setting with the model quantized to 4-bit. We mainly adopt *6-16* as it achieves the overall best results, balancing the strength of augmentations and the stability in training.

Table 9: Ablation studies of CQ-C with different precision sets on CIFAR-100.

Network	**Precision Set**	Fine-tune Acc. (**FP**)		Fine-tune Acc. (**4-bit**)	
		10% labels	**1%** labels	**10%** labels	**1%** labels
ResNet-18	4-16	61.82	42.95	60.4	41.91
	6-16	61.75	43.8	60.12	42.59
	8-16	61.34	43.15	59.91	41.66
ResNet-34	4-16	63.42	46.95	62.3	45.45
	6-16	63.58	48.05	61.47	45.75
	8-16	63.8	48.12	61.85	45.81
ResNet-74	4-16	53.04	31.63	52.08	30.30
	6-16	52.52	31.39	51.12	29.70
	8-16	53.24	31.26	51.57	30.20
ResNet-110	4-16	53.37	31.96	52.03	31.25
	6-16	54.47	33.17	52.28	32.66
	8-16	53.17	33.04	52.18	31.14
MobileNetV2	4-16	51.63	25.54	50.16	22.61
	6-16	51.59	26.12	49.82	20.82
	8-16	51.33	25.1	48.99	19.7