Reproducibility report

Sangbeom Jeong
Department of Electrical and Information Engineering
Seoul National University of Science and Technology
sangbeom@seoultech.ac.kr

Seongjun O
Department of Electrical and Information Engineering
Seoul National University of Science and Technology
seongjun_o@seoultech.ac.kr

Junghyeok Lee
Department of Electrical and Information Engineering
Seoul National University of Science and Technology
ljhy20k@seoultech.ac.kr

Abstract

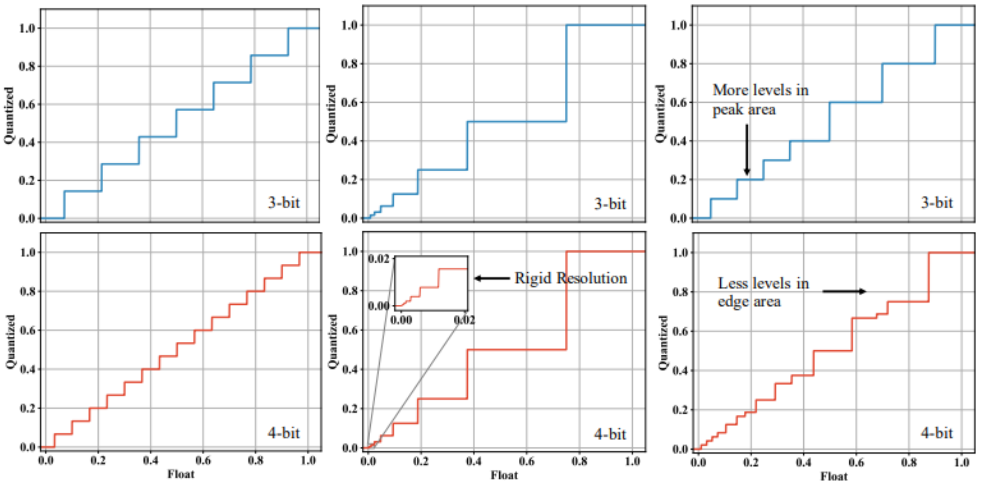
This study aims to reproduce the main results of the paper "Additive Powers-of-Two Quantization: An Efficient Non-uniform Discretization for Neural Networks" (ICLR 2020). We focus on applying the additive powers-of-two (APoT) quantization technique to ResNet architectures and evaluating its performance on the CIFAR-10, CIFAR-100, and ImageNet datasets. The original paper claims that APoT quantization achieves efficient non-uniform quantization by exploiting the bell-shaped and long-tailed distribution characteristics of weights and activations, providing competitive accuracy compared to full-precision implementations with higher computational efficiency. We implement the core elements of APoT in a PyTorch environment and conduct experiments using ResNet-20/56 architectures for CIFAR-10/100 and ResNet-18 for ImageNet. The experimental results on CIFAR-10 show that applying APoT results in less than a 1% drop in accuracy for 4, 3, and 2-bit quantization, replicating the trends observed in the original paper. However, there is a slight discrepancy in the 5-bit quantization results for the ResNet-18 model on ImageNet. Additional experiments on CIFAR-100 reveal a minor accuracy drop for 4-bit quantization, while a more significant drop is observed for 3-bit and 2-bit quantization. These discrepancies may be attributed to hardware differences and variations in training settings. Despite these limitations, our study confirms the effectiveness of APoT quantization in achieving competitive accuracy with reduced bit-widths, highlighting its potential for efficient neural network compression. Our code can be found at the following :https://github.com/SangbeomJeong/Reproducibility-Project

1 Introduction

In this article, we replicate the results of additive powers-of-two (APoT) quantization[Li et al., 2020] for bell-shaped and long-tailed distributions of weights and activations in neural networks. APoT quantization is a non-uniform quantization method recognized as the state-of-the-art quantization technique as of 2020. Previous studies [Cai et al., 2017, Gong et al., 2019] have predominantly focused on uniform quantization methods due to their simplicity and ease of implementation in hardware. However, these methods do not consider the non-uniform distribution of weights and activations in deep neural networks (DNNs), which typically follow a bell-shaped and long-tailed distribution. This mismatch leads to suboptimal performance, as a large proportion of weights are concentrated around the mean, with a few outliers having relatively high magnitudes. To solve this problem, the APoT quantization scheme exploits the advantages of non-uniform quantization while maintaining computational efficiency. APoT quantization utilizes a sum of multiple powers-of-two terms to create quantization levels that adapt well to the bell-shaped distribution of weights, offering high resolution near the mean and sufficient representation for outliers. This approach significantly reduces computational overhead, achieving approximately 2× speed-up in multiplication operations compared to uniform quantization. Additionally, we propose a reparameterized clipping function (RCF) to optimize the clipping threshold more accurately, enhancing the quantization process. By introducing weight normalization, we ensure stable and consistent clipping and projection of weights during the forward pass. Therefore, as seen in Figure 1, APoT quantization allows for more reasonable resolution assignment and is not affected by the constraints of resolution.

This article aims to enhance the transparency and reliability of scientific research by verifying the claims of the original paper and reproducing its results. Additionally, this study seeks to advance scientific progress by adding new research insights and contributions to the literature. By improving investment incentives for reproducibility studies, we aim to

33rd Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.



(a) Uniform Quantization (b) Power-of-Two Quantization (c) Additive Power-of-Two Quantization

Figure 1: **Results of the APoT technique when quantizing unsigned data to 3-bit or 4-bit (α = 1.0) using three different quantization levels.**

enable the research community to make higher-quality scientific contributions. Through these efforts, we emphasize the importance of reproducibility as a critical scientific methodology and contribute to enhancing the accuracy and reliability of research.

2 Scope of reproducibility

This section provides a detailed explanation of the main contributions and techniques discussed in the original paper.

Powers-of-Two Quantization

To improve both quantization resolution and computational efficiency, APoT is based on the powers-of-two quantization technique, which restricts the quantization levels to powers-of-two values or zero [Miyashita et al., 2016, Zhou et al., 2016]. powers-of-two quantization constrains the quantization levels to powers-of-two values or zero. This is formally defined as:

$$Q_p(\alpha, b) = \alpha \times \{0, \pm 2^{-(2^{b-1}-1)+1}, \pm 2^{-(2^{b-1}-1)+2}, \dots, \pm 2^{-1}, \pm 1\}$$

where α is a scaling factor, and b represents the bit-width. Multiplications involving a power-of-two number 2^x and another operand r are efficiently implemented using bit-wise shift operations as follows:

$$2^x r = \begin{cases} r & \text{if } x = 0 \\ r \ll x & \text{if } x > 0 \\ r \gg x & \text{if } x < 0 \end{cases}$$

where $\ll$ and $\gg$ denote left and right shift operations, respectively. Despite its efficiency, PoT quantization suffers from rigid resolution issues, where increasing the bit-width b only enhances resolution near zero, resulting in large projection errors for higher values.

Additive Powers-of-Two (APoT) Quantization

APoT Quantization is an advanced quantization technique aimed at achieving efficient non-uniform quantization for neural networks. This method leverages the distribution characteristics of neural network weights and activations, which often follow a bell-shaped and long-tailed distribution, to improve both quantization resolution and computational efficiency. APoT quantization addresses the limitations of PoT by summing multiple PoT terms to create more flexible and precise quantization levels. The quantization levels in APoT are defined as follows:

$$Q_a(\alpha, kn) = \gamma \times \left\{ \sum_{i=0}^{n-1} p_i \right\} \quad \text{where} \quad p_i \in \left\{ 0, \frac{1}{2^i}, \frac{1}{2^{i+n}}, \dots, \frac{1}{2^{i+(2k-2)n}} \right\}$$

Here, γ is a scaling coefficient ensuring the maximum level in Q_a is α. The variable k is the base bit-width for each additive term, and n is the number of additive terms. Given the bit-width b and base bit-width k, n is calculated as $n = \frac{b}{k}$. Consequently, there are $2^{kn} = 2^b$ quantization levels.

The quantization levels are given by $\gamma \times (p_0 + p_1)$ for all combinations of p_0 and p_1, resulting in 16 unique combinations. This approach allocates quantization levels more judiciously, especially in the central region, and increases the resolution for higher value ranges, mitigating the rigid resolution issue inherent in PoT quantization.

In this section, we itemize the main contributions claimed by APoT:

- The APoT paper demonstrated that the method achieves competitive accuracy comparable to full-precision implementations with higher computational efficiency.
- Specifically, the 4-bit quantized ResNet-50 on ImageNet achieved 76.6% Top-1 and 93.1% Top-5 accuracy.
- Compared to uniform quantization methods, the proposed algorithm reduces computational costs by 22%, proving its hardware-friendly nature.

3 Methodology

3.1 Model descriptions

In this study, we experimented with the APoT quantization technique using ResNet-18, ResNet-20, and ResNet-56 models. ResNet-18 is a model used on the ImageNet dataset, with 18 layers. This model has 11.7M parameters and takes 224x224 size images as input to classify them into 1000 classes. ResNet-20 and ResNet-56 are models used on the CIFAR-10 and CIFAR-100 datasets. ResNet-20 has 20 layers and approximately 0.27M parameters. ResNet-56 has 56 layers and approximately 0.85M parameters. These models take 32x32 size images as input and classify them into 10 or 100 classes. Initially, quantization experiments were conducted using pretrained weights provided by the authors of the paper, but performance was lower in all conditions. So, we conducted an experiment by relearning the baseline from the beginning and securing pretrained weights.

3.2 Datasets

The experiments were conducted using the CIFAR-10, CIFAR-100, and ImageNet datasets. The CIFAR-10 dataset consists of 60,000 color images with a size of 32x32 pixels, divided into 10 classes (airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck). Each class contains 6,000 images, with 50,000 images used for training and 10,000 images used for testing. The CIFAR-100 dataset has the same image size and quantity as CIFAR-10 but is composed of 100 classes. Each class has 600 images, with 500 images used for training and 100 images used for testing. ImageNet is a large-scale dataset consisting of 1000 classes. The training data comprises approximately 1.2 million images, and the validation data consists of 50,000 images.

3.3 Hyperparameters

For experiments on the CIFAR-10 and CIFAR-100 datasets, the number of epochs was set to 300, and the batch size was set to 128. stochastic gradient descent (SGD) was used as the optimizer, with a weight decay of either 5e-4 or 1e-4. The learning rate was set to 4e-2 and adjusted using the MultiStepLR scheduler. For the ImageNet dataset, the number of epochs was set to 120, and the batch size was set to 256.

3.4 Experimental setup and code

All experiments were implemented using the PyTorch framework. The experimental code is available in the GitHub repository (https://github.com/SangbeomJeong/Reproducibility-Project.git). The performance of the models was evaluated based on the accuracy achieved on the test. The experiments were conducted following these steps:
1.Train the baseline model with full-precision bit-width.
2.Load the trained weights and apply APoT quantization.
3.Evaluate the performance of the quantized models for various bit-widths (4/3/2-bit).

3.5 Computational requirements

The experiments were performed using two NVIDIA GeForce RTX 2080 Ti GPUs. PyTorch version was 2.2.1, and CUDA version was 12.1. Experiments on the CIFAR-10 and CIFAR-100 datasets took approximately 3 hours per experiment (300 epochs) using a batch size of 128. Experiments on the ImageNet dataset took an average of 80 minutes per epoch using a batch size of 256. It took approximately 160 hours to train a total of 120 epochs.

4 Results

4.1 Results reproducing original paper

In this section, we present results that support one of the main contributions of Section 2, demonstrating that APoT quantization achieves accuracy comparable to full-precision implementations with significantly higher computational efficiency.

4.1.1 APoT Quantization Method Reproducibility for CIFAR-10

Table 1 summarizes the accuracy and accuracy drop for both the original paper and our reproducibility experiments. For ResNet20, the original paper reported that 4-bit and 3-bit quantization slightly exceeded the baseline accuracy, with only the 2-bit quantization showing a small drop. In contrast, our reproducibility results showed a consistent decrease in accuracy across all bit-precision settings. Specifically, our 4-bit quantization resulted in a 0.34% drop, 3-bit quantization had a 0.37% drop, and 2-bit quantization had a significant 1.88% drop. For ResNet56, the original paper indicated performance gains for both 4-bit and 3-bit quantization. However, our results showed that the 4-bit quantization resulted in a 0.09% drop, the 3-bit quantization had a 0.48% drop, and the 2-bit quantization showed a 1.19% drop. Therefore, while the original paper demonstrated the benefits of the APoT quantization technique, our reproducibility study revealed a drop in accuracy across all settings.

Architecture	Bit-precision	Original Paper		Reproducibility (Ours)	
		Accuracy (%)	Acc. Drop (%)	Accuracy (%)	Acc. Drop (%)
ResNet20	Baseline (FP32)	91.6	-	92.90	-
	4bit	92.3	-0.7	92.56	0.34
	3bit	92.2	-0.6	92.53	0.37
	2bit	91.0	0.6	91.02	1.88
ResNet56	Baseline (FP32)	93.2	-	93.94	-
	4bit	94.0	-1.8	93.85	0.09
	3bit	93.9	-0.7	93.46	0.48
	2bit	92.9	0.3	92.75	1.19

Table 1: Comparison of accuracy and accuracy drop between the original paper results and our reproducibility results for ResNet20 and ResNet56.

4.1.2 APoT Quantization Method Reproducibility for ImageNet

The accuracy of applying the APoT quantization method to ResNet18 on ImageNet can be seen in Table 2. While the original paper reported improved accuracy with 5-bit precision compared to the baseline, our reproducibility experiments showed a decrease in accuracy, with a drop of 0.81%.

Architecture	Bit-precision	Original Paper		Reproducibility (Ours)	
		Accuracy (%)	Acc. Drop (%)	Accuracy (%)	Acc. Drop (%)
ResNet18	Baseline (FP32)	70.2	-	70.2	-
	5bit	70.9	-0.7	69.39	0.81

Table 2: Comparison of accuracy and accuracy drop between the original paper results and our reproducibility results for ResNet18.

4.2 Results beyond original paper

To verify the effectiveness of the APoT quantization method on different datasets, we conduct an ablation study and perform hyperparameter tuning. First, we carry out experiments on CIFAR-100 can be seen in Table 3, and the results of the second experiment, which uses MultiStepLR in this paper to adaptively apply the learning rate but replaces it with CosineAnnealingLR to improve performance, can be seen in Table 4.

4.2.1 ResNet models on CIFAR-100

Table 3 summarizes the accuracy and accuracy drop for both the original paper and our reproducibility experiments on CIFAR-100. For ResNet20, the baseline accuracy using full precision (i.e., FP32) was 69.09%. When applying the APoT quantization method, the 4-bit quantization resulted in a 0.53% drop, the 3-bit quantization had a 1.53% drop, and the 2-bit quantization showed a significant 5.17% drop. additionally, The 2-bit quantization of ResNet56 also showed a significant drop of 3.28%.

Architecture	Bit-precision	Accuracy (%)	Acc. Drop (%)
ResNet20	Baseline (FP32)	69.09	-
	4bit	68.56	0.53
	3bit	67.56	1.53
	2bit	63.92	5.17
ResNet56	Baseline (FP32)	72.92	-
	4bit	72.19	0.73
	3bit	71.47	1.45
	2bit	69.64	3.28

Table 3: Experimental Results of ResNet Models Applying APoT Quantization Method on CIFAR-100.

4.2.2 ResNet models on CIFAR-10 with CosineAnnealingLR

In an attempt to further improve the previously reproduced performance, we will change the lr scheduler to CosineAnnealingLR and check the results. All environments except the lr scheduler are the same as the existing CIFAR-10 reproducibility environment. Applying CosineAnnealingLR to improve accuracy, as shown in Table 4, resulted in lower accuracy compared to the conventional MultiStepLR used by APoT. Therefore, our ablation study confirms that the existing method is more effective.

Architecture	Bit-precision	Accuracy (%)	Acc. Drop (%)
ResNet20	Baseline (FP32)	92.94	-
	4bit	92.6	0.34
	3bit	92.17	0.77
	2bit	90.67	2.27

Table 4: Experimental Results of ResNet Model Applying APoT Quantization Method on CIFAR-10 with the Learning Rate Scheduler Changed to CosineAnnealingLR.

5 Discussion

The experimental results of this study partially support the effectiveness of the APoT quantization technique. On the CIFAR-10 dataset, the reproduced models exhibited slightly lower performance compared to the original paper, but still maintained high accuracy even at low bit-widths. For the ImageNet dataset, only 5-bit quantization was reproduced, but the achieved accuracy was similar to that reported in the original paper. As a result of conducting experiments with CosineAnnealingLR, 4-bit quantization recorded the same accuracy drop as MultiStepLR. However, 3/2bit quantization showed a much larger accuracy drop than MultiStepLR. Through this, we concluded that the quantization method proposed in this paper is more effective when using MultiStepLR. This study has several limitations. Firstly, due to hardware constraints, the experiments on ImageNet were not conducted extensively enough to evaluate the performance at various bit-widths. Moreover, not all experimental settings presented in the original paper were perfectly reproduced, which can be attributed to differences in implementation, experimental environment, and limitations in hyperparameter search. Despite these limitations, this study is meaningful in that it confirms the effectiveness of the APoT quantization technique and presents new findings through additional experiments. The results on the CIFAR-100 dataset suggest the generalizability of APoT quantization, and the possibility of performance improvement by changing the learning rate scheduler was also verified. Future research should apply APoT quantization to a wider range of datasets and models, and fine-tune hyperparameters such as quantization bit-width and clipping range. Furthermore, it would be worthwhile to quantitatively analyze the effects of quantization on model compression and inference speed. In conclusion, this study verified the effectiveness of APoT quantization through reproduction and provided new insights through additional experiments. Quantization is one of the important methods for improving the efficiency of deep learning models, and it is expected to make further progress through the development of techniques such as APoT.

5.1 What was easy

The publicly available code provided by the authors was well-implemented, allowing us to set up the experimental environment and conduct reproducibility experiments with ease. The core idea of APoT for configuring quantization levels was intuitive, facilitating easy understanding and implementation. Specifically, the quantization of activations and weights, as well as convolution operations, were effectively implemented in PyTorch code, making them readily usable. Additionally, the detailed methodological explanations in the paper greatly assisted in understanding and replicating the experiments. Initializing the quantization experiments using pre-trained weights was also relatively straightforward.

5.2 What was difficult

Securing computational resources and the long training times required for the experiments posed significant challenges. Specifically, GPU memory limitations for training on ImageNet made it difficult to experiment with deeper models such as ResNet-50. Additionally, setting up the experimental environment for CIFAR-100 involved some trial and error. Variations in hardware configuration led to slight deviations in performance metrics. The low-bit (2-bit) quantization experiments required hyperparameter tuning to achieve reasonable accuracy, demanding substantial computational resources and time. While the main trends of the APoT paper were confirmed for CIFAR-10, conducting larger-scale experiments was constrained by resource limitations.

References

Zhaowei Cai, Xiaodong He, Jian Sun, and Nuno Vasconcelos. Deep learning with low precision by half-wave gaussian quantization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5918–5926, 2017.

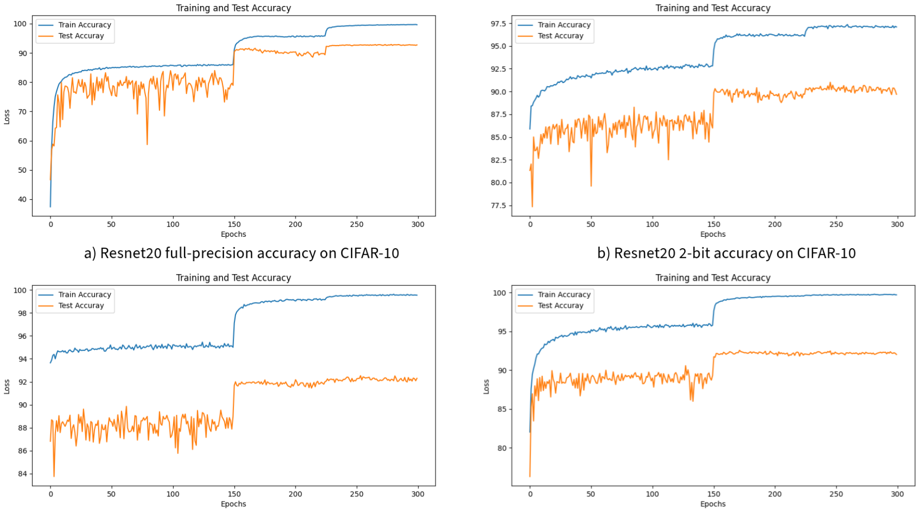
Ruihao Gong, Xianglong Liu, Shenghu Jiang, Tianxiang Li, Peng Hu, Jiazhen Lin, Fengwei Yu, and Junjie Yan. Differentiable soft quantization: Bridging full-precision and low-bit neural networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4852–4861, 2019.

Yuhang Li, Xin Dong, and Wei Wang. Additive powers-of-two quantization: An efficient non-uniform discretization for neural networks. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=BkgXT24tDS`.

Daisuke Miyashita, Edward H Lee, and Boris Murmann. Convolutional neural networks using logarithmic data representation. *arXiv preprint arXiv:1603.01025*, 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.

Appendix A



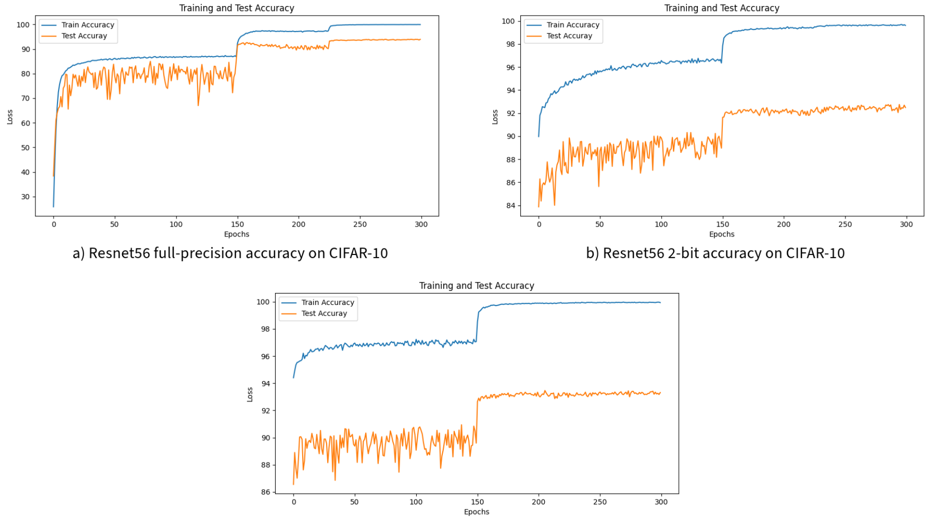
a) Resnet20 full-precision accuracy on CIFAR-10

b) Resnet20 2-bit accuracy on CIFAR-10

c) Resnet20 3-bit accuracy on CIFAR-10

d) Resnet20 4-bit accuracy on CIFAR-10

Figure 2: **Accuracy Results of the ResNet-20 Applying the APoT Quantization Method on CIFAR-10.**



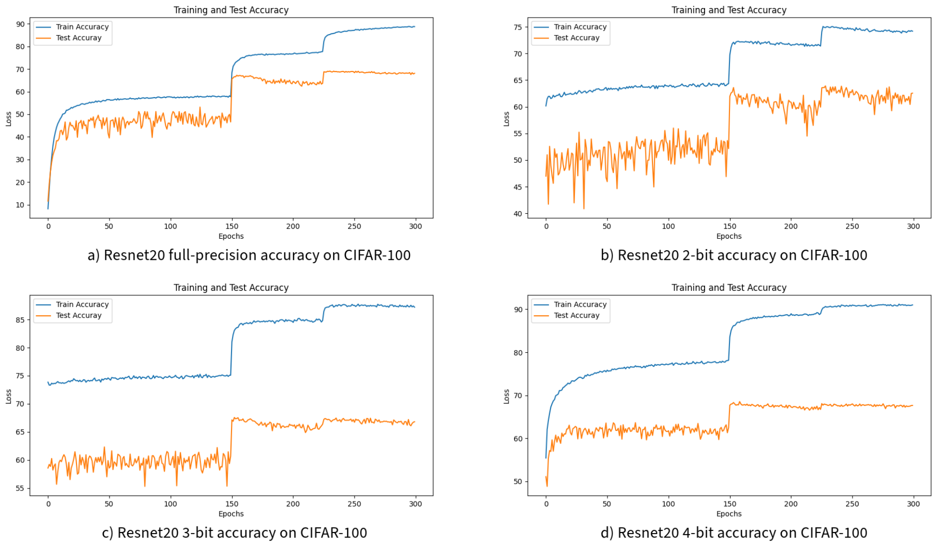
a) Resnet56 full-precision accuracy on CIFAR-10

b) Resnet56 2-bit accuracy on CIFAR-10

c) Resnet56 3-bit accuracy on CIFAR-10

Figure 3: **Accuracy Results of the ResNet-56 Applying the APoT Quantization Method on CIFAR-10.**

Appendix B



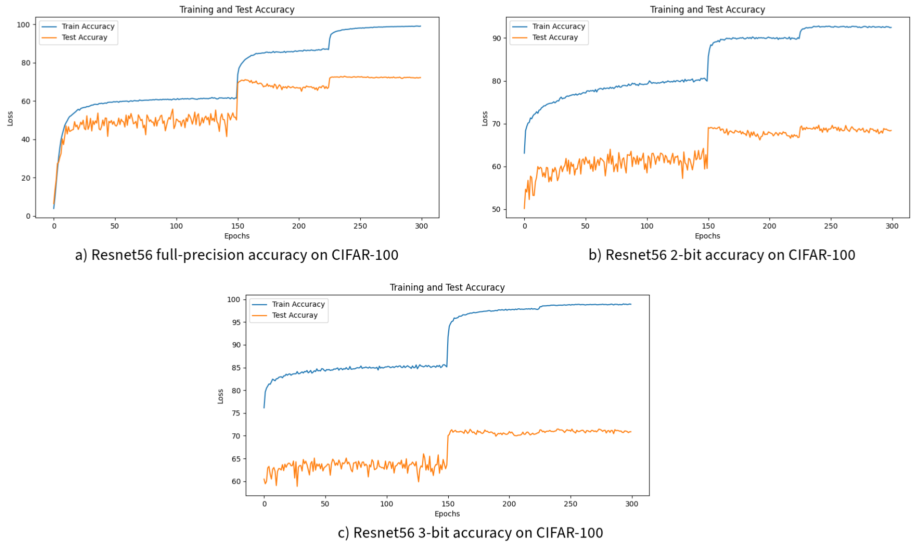
a) Resnet20 full-precision accuracy on CIFAR-100

b) Resnet20 2-bit accuracy on CIFAR-100

c) Resnet20 3-bit accuracy on CIFAR-100

d) Resnet20 4-bit accuracy on CIFAR-100

Figure 4: **Accuracy Results of the ResNet-20 Applying the APoT Quantization Method on CIFAR-100.**



a) Resnet56 full-precision accuracy on CIFAR-100

b) Resnet56 2-bit accuracy on CIFAR-100

c) Resnet56 3-bit accuracy on CIFAR-100

Figure 5: **Accuracy Results of the ResNet-56 Applying the APoT Quantization Method on CIFAR-100.**