Appendix for "Estimator Meets Equilibrium Perspective: A Rectified Straight Through Estimator for Binary Neural Networks Training"

Xiao-Ming Wu[1], Dian Zheng[1], Zuhao Liu[1], Wei-Shi Zheng[1,2,3,4*]
[1]School of Computer Science and Engineering, Sun Yat-sen University, China, [2]Pengcheng Lab, China,
[3]Guangdong Province Key Laboratory of Information Security Technology, China,
[4]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
{wuxm65, zhengd35, liuzh327}@mail2.sysu.edu.cn, wszheng@ieee.org

1. Exploration of binary ViT

Vision Transformer [2] has received widespread attention in computer vision community due to its amazing performance in recent years. To this end, we further explore to binarize ViT using ReSTE. Similar to the binarization in convolutional neural network, we binarize all the attention and MLP layers expect for the first layer and the last classification head. We follow the typical setting in ViT training. For specific, we use ImageNet ILSVRC-2012 [1] dataset and train the model from scratch. RandomCrop, RandomHorizontalFlip and Normalize strategies are applied for data pre-processing. We use AdamW optimizer and set the beginning learning rate equals to 0.001. We apply Cosine learning late descent schedule in training. Cross entropy is adopted as the loss function. As for the hyper-parameter o_{end}, we set $o_{\text{end}} = 3$. We use ViT tiny as backbone and compare ReSTE with the baseline model DoReFa-Net[5]. The results are shown in Table 1.

Backbone	Estimator	W/A	Top-1(%)	Top-5(%)
ViT tiny	FP	32/32	64.50	85.14
	DoReFa-Net[5]	1/32	53.05	76.83
	ReSTE (ours)	1/32	**56.53**	**79.86**

Table 1: Performance in ImageNet dataset with ViT as backbone. FP is the full-precision version of the backbone. W/A is the bit width of weights or activations. Best results are shown in black bold font.

From the table we can observe that ReSTE has excellent performance, with about 3.48% and 3.03% improvement over the baseline model DoReFa-Net[5] at Top-1 and Top-5 accuracy, which further validates the effectiveness and wide applicability of our method. In addition, it can be seen that the desirable value $o_{\text{end}} = 3$ in binary convolutional neural network is also applicable in binary ViT architecture, which shows the robustness of our estimator.

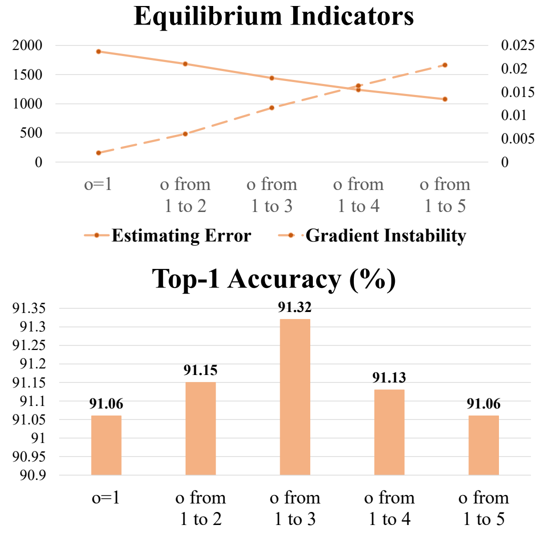
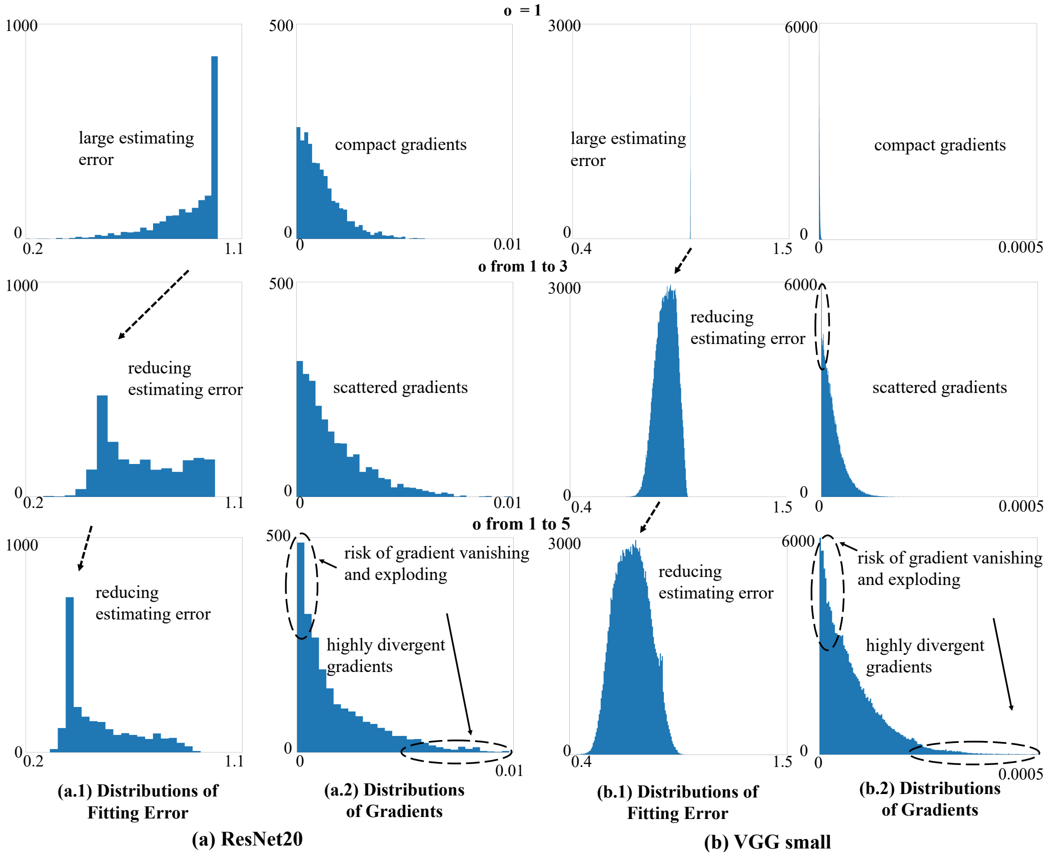


Figure 1: Illustrations of the estimating error indicators (above), gradient instability indicators (above) and the Top-1 accuracy (below) with different scales of o_{end} at the setting of 1W/32A.

Nevertheless, we can see that binary ViT has a nonnegligible performance degradation comparing with the full-precision model. The reason is that attention needs discriminate features to produce highly differentiable attention map, but binarization reduces the representative ability of features. Exploring more suitable attention block specifically for binarization is our future work.

[*] denotes the corresponding author.



(a.1) Distributions of Fitting Error

(a.2) Distributions of Gradients

(b.1) Distributions of Fitting Error

(b.2) Distributions of Gradients

(a) ResNet20

(b) VGG small

Figure 2: Illustrations of the distributions of the fitting error (left) and the gradients (right) with ResNet20 and VGG small as backbone.

2. More Analysis about Equilibrium Perspective

To further demonstrate our equilibrium perspective, we conduct some additional analysis experiments to reveal it. We firstly test the estimating error, gradient stability and the model performance with different scales of o_{end} at the setting of 1W/32A (experiments at the setting of 1W/1A have been shown in the main body of our paper). The experiments use ResNet-20 as backbone and test on CIFAR-10[3] dataset.

The results are shown in Fig. 1. From the figures we can get the similar conclusions as the experiments at the setting of 1W/1A, which is shown in the main body of our paper. We can see that the estimating error becomes smaller and smaller and the gradient instability becomes bigger and bigger with o_{end} increasing. This observation shows that we can not reduce the estimating error with no limit since the gradient stability will decline along with. In addition, with the change of o_{end}, the model performance increases first and then decreases, implying that the large gradient instability will harm the model performance, which validates our claim.

In addition, it can also be seen from the figures that ReSTE can flexibly adjust the degree of the estimating error and the gradient stability by easily changing the hyperparameter o_{end}, which implies the superiority of our method. Moreover, the desirable degrees of equilibrium, i.e., the desirable o_{end} to produce high performance, is same with the settings of 1W/1A, which is shown in the main body of our paper, further showing the robustness and wide applicability of ReSTE.

In addition, we also show more visualizations about the distributions of the fitting error and the gradients with ResNet-20 and VGG-small as backbone. The experiments are conducted in CIFAR-10 [3] dataset at the setting of 1W/1A. The results are shown in Fig. 2. The conclusions are similar to the experiments with ResNet-18 as backbone, which is shown in the main body of our paper. The peak values of the estimating error distribution become smaller, but the gradients become more divergent, which harms the model training and increases the risk of gradient vanishing or exploding. This visualization further demonstrate the equilibrium phenomenon and validates the necessity to balance it.

3. Model Design Experiments

In model design experiments, we firstly conduct experiments about two different adjusting strategies of o in ReSTE, the fixed strategy and the progressive strategy, which is proposed in [4]. For fixed strategy, we test different scales of o, and about the progressive strategy we try different values of o_{end}. Experiments are conducted with ResNet-20 as backbone in CIFAR-10 [3] dataset, at the setting of 1W/1A. All the results are demonstrated in Table 3. From the table we can observe that the progressive strategy is better than the fixed strategy, since the progressive strategy allows sufficient updating at the beginning and accurate gradients at the end of the training. This can also be explained by the equilibrium perspective. If we use fixed strategy, we have divergent gradients at the beginning of the training, which leads to wrong update directions at the very beginning and dramatically harms the model training.

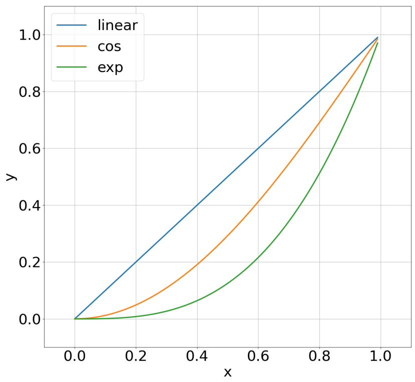


Figure 3: Visualizations of different tuning ways of o in the progressive strategy.

Strategy		Acc (%)	Strategy		Acc (%)
Progressive	$o_{\text{end}} = 1$	85.18	Fixed	$o = 1$	85.18
	$o_{\text{end}} = 2$	86.41		$o = 2$	85.80
	$o_{\text{end}} = 3$	86.75		$o = 3$	83.19
	$o_{\text{end}} = 4$	86.45		$o = 4$	82.04

Table 2: Model performance of different adjusting strategies of o with ResNet-20 as backbone.

In addition, we also test different tuning ways in the progressive strategy. In detail, we test cos, exp and linear tuning ways, whose visualizations are shown in Fig. 3. Experiments are conducted with ResNet-20 as backbone in CIFAR-10 [3] dataset at the setting of 1W/1A. All the results are demonstrated in Table 3. From the table we can observe that different tuning ways in the progressive strategy have similar performance, showing the robustness of our method. We simply use cos tuning way in all the experiments in our paper.

Strategies	Acc(%)
Linear	86.80
Cosine	86.75
Power	86.59

Table 3: Model performance of different tuning ways in the progressive strategy.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[3] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Technical report, University of Toronto*, 2009.

[4] Haotong Qin, Ruihao Gong, Xianglong Liu, Mingzhu Shen, Ziran Wei, Fengwei Yu, and Jingkuan Song. Forward and backward information retention for accurate binary neural networks. In *CVPR*, 2020.

[5] 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*, 2016.