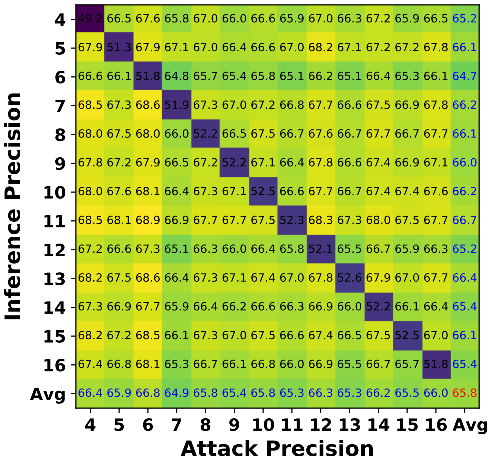
Supplementary Materials of Double-Win Quant: Aggressively Winning Robustness of Quantized Deep Neural Networks via Random Precision Training and Inference

Yonggan Fu [1] Qixuan Yu [1] Meng Li [2] Vikas Chandra [2] Yingyan Lin [1]



	4	5	6	7	8	9	10	11	12	13	14	15	16	Avg
4	49.2	66.5	67.6	65.8	67.0	66.0	66.6	65.9	67.0	66.3	67.2	65.9	66.5	65.2
5	67.9	51.3	67.9	67.1	67.0	66.4	66.6	67.0	68.2	67.1	67.2	67.2	67.8	66.1
6	66.6	66.1	51.8	64.8	65.7	65.4	65.8	65.1	66.2	65.1	66.4	65.3	66.1	64.7
7	68.5	67.3	68.6	51.9	67.3	67.0	67.2	66.8	67.7	66.6	67.5	66.9	67.8	66.2
8	68.0	67.5	68.0	66.0	52.2	66.5	67.5	66.7	67.6	66.7	67.7	66.7	67.7	66.1
9	67.8	67.2	67.9	66.5	67.2	52.2	67.1	66.4	67.8	66.6	67.4	66.9	67.1	66.0
10	68.0	67.6	68.1	66.4	67.3	67.1	52.5	66.6	67.7	66.7	67.4	67.4	67.6	66.2
11	68.5	68.1	68.9	66.9	67.7	67.7	67.5	52.3	68.3	67.3	68.0	67.5	67.7	66.7
12	67.2	66.6	67.3	65.1	66.3	66.0	66.4	65.8	52.1	65.5	66.7	65.9	66.3	65.2
13	68.2	67.5	68.6	66.4	67.3	67.1	67.4	67.0	67.8	52.6	67.9	67.0	67.7	66.4
14	67.3	66.9	67.7	65.9	66.4	66.2	66.6	66.3	66.9	66.0	52.2	66.1	66.4	65.4
15	68.2	67.2	68.5	66.1	67.3	67.0	67.5	66.6	67.4	66.5	67.5	52.5	67.0	66.1
16	67.4	66.8	68.1	65.3	66.7	66.1	66.8	66.0	66.9	65.5	66.7	65.7	51.8	65.4
Avg	66.4	65.9	66.8	64.9	65.8	65.4	65.8	65.3	66.3	65.3	66.2	65.5	66.0	65.8

Figure 1. Visualizing the transferability of adversarial attacks between different precisions of WideResNet32 trained using the enhanced DWQ, where we annotate the robust accuracy under the PGD-20 attack, and provide both the average accuracy of each row and column (annotated in blue) and the overall average accuracy (annotated in red), the latter of which is also the robust accuracy if RPI is adopted.

1. Visualizing the transferability heatmap for WideResNet32 trained by enhanced DWQ

Experiment settings. Fig. 1 visualizes the transferability of PGD-20 attacks between different precisions on WideResNet32 trained using the enhanced DWQ with a training/inference precision set of 4∼16-bit, where we also annotate both the overall average accuracy (in red) and the average precision of each row and column (in blue) for measuring the robustness of each precision choice.

Results and analysis. In addition to the consistently poor transferability of generic quantized DNNs as identified in Fig. 1 of the main content, we also observe that (1) the enhanced DWQ equipped with RPT and switchable BNs notably enlarges the robust accuracy gaps between different precisions especially under higher precisions, i.e., **the enhanced DWQ further increases the difficulty of transferring adversarial attacks between different precisions** as compared with vanilla DWQ shown in Fig. 1 of the main content, and (2) the average robust accuracies under each attack precision (i.e., each column) are close to each other (within 2%), i.e., **there's no notable bottleneck precision that can be utilized by the attackers**. This set of experiment indicates that although both attackers and defenders can adjust the sampling strategies of their attack/inference precisions for better attack/defense in their competition, DWQ is generally effective and applicable, and our assumption of the random precision strategy in Sec. 4.1 of the main content can fairly evaluate the effectiveness of DWQ.

[1] Department of Electrical and Computer Engineering, Rice University [2] Facebook Inc. Correspondence to: Yingyan Lin <yingyan.lin@rice.edu>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).