ODG-Q: Robust Quantization via Online Domain Generalization

Chaofan Tao, Ngai Wong
The University of Hong Kong
Email: cftao@eee.hku.hk, nwong@eee.hku.hk

Abstract—Quantizing neural networks to low-bitwidth is important for model deployment on resource-limited edge hardware. Although a quantized network has a smaller model size and memory footprint, it is fragile to adversarial attacks. However, few methods study the robustness and training efficiency of quantized networks. To this end, we propose a new method by recasting robust quantization as an online domain generalization problem, termed ODG-Q, which generates diverse adversarial data at a low cost during training. ODG-Q consistently outperforms existing works against various adversarial attacks. For example, on CIFAR-10 dataset, ODG-Q achieves 49.2% average improvements under five common white-box attacks and 21.7% average improvements under five common black-box attacks, with a training cost similar to that of natural training (viz. without adversaries). To our best knowledge, this work is the first work that trains both quantized and binary neural networks on ImageNet that consistently improves robustness under different attacks. We also provide a theoretical insight of ODG-Q that accounts for the bound of model risk on attacked data.

I. INTRODUCTION

Despite the impressive performance of deep neural networks (DNNs) in various tasks [3], [22], [2], they are vulnerable to adversarial attacks [6]. A well-trained DNN can be easily fooled to misclassify by only a small (but carefully computed) perturbation on the input data that is even not noticeable to the human eyes. Such adversarial attacks can lead to serious security and safety concerns such as autonomous driving, mobile robots, face recognition, etc. On the other hand, quantization [1], [26], [23] has become an essential step for DNN deployment, especially on resource-limited edge devices. Quantization converts the data format of weights and activation from floating-point (FP) to low-bitwidth integer (INT), thereby significantly reducing the computation and storage budgets.

Although quantization (or binarization in the extreme case) can largely reduce model deployment cost, a natural question arises: are quantized models robust enough to defend against adversarial attacks? According to [12], models are vulnerable under adversarial attacks when they are quantized to low bitwidths, e.g., ≤ 4 bits. When quantized to 1-bit, one gets binary neural networks [16], [13] that are extremely hardware-friendly but their robustness against attacks, especially for large-scale problems, remains underexplored. In addition, to improve model robustness, most previous defense methods [15], [5], [30] based on the FP models fall short to balance efficiency and robustness. For example, training with augmented adversarial samples offline can improve robustness but typically incurs 3× ~ 30× overhead vs natural training [20], i.e., training on clean data without defense consideration. For example, Ref. [28] trains a robust classifier on ImageNet using 128 V100 cards at around 31× computational cost of natural training. Also, adversarial training often requires extra storage for adversarial samples, e.g., the PGD-\( K \) training [15] requires an outer loop for iterating the mini-batch data similar to natural training, plus an inner loop to generate \( K \)-step adversarial samples. Regularization-based method [19] enhances robustness by regularizing weights and/or activation, which does not require a huge training cost but their performance is often limited especially for large-scale datasets. AdvFree [20] uses 4 P100 GPUs and 2 days to train a FP model on ImageNet. AdvFree improves efficiency by recycling gradients that update the weights to generate adversarial perturbations. However, it still trains with each mini-batch for multiple loops per epoch to generate adversarial updates.

We argue that the holy grail is to generate diverse adversarial samples at a very low cost. Our main innovation is to recast the learning of adversarial robustness into a domain generalization problem. The natural samples are viewed as a subset of source data. Unlike previous defense methods [6], [15], [30] that only consider local adversarial perturbations generated from each sample itself in each batch, we argue that global adversarial perturbations in different batches are also important, which help the model to generate diverse adversarial samples. By constructing a set of adversarial perturbations and updating the subset of perturbations in turn during each iteration like a
roulette, we obtain a bunch of adversarial samples from multiple source domains which have different times of adversarial updating. The adversarial perturbations are generated online so that different samples arrive in a batch-sequential order without an inner loop, thereby reducing the training time sharply and obviating the need to store adversarial samples (cf. Fig. 1).

To sum up, the contributions of this paper are threefold. 1) We recast the learning of adversarial robustness as a domain generalization problem. By generating adversarial source data from multiple domains online during training, and enforcing high-level feature alignment, the proposed algorithm achieves state-of-the-art robustness with a runtime close to that of natural training. 2) Inspired by the theory of domain generalization, we also analytically prove that the risk on target data (attacked samples on the inference model) can be bounded with the proposed approach. 3) Extensive experiments on MNIST, CIFAR-10 and ImageNet datasets demonstrate the effectiveness of the proposed OD-G, under various adversarial attacks in both white-box and black-box settings.

II. RELATED WORK AND PRELIMINARIES

A. Objective of Robustness Learning

Definition 1. Given a natural dataset \((x, y)\) \((x: \text{data}, y: \text{label})\) sampled from a domain \(D\) and a task loss function \(L_{\Theta}(\cdot, \cdot)\), the goal of learning a robust model is to learn parameters \(\Theta\) that maximize the performance under any perturbation \(p\) (within an \(\epsilon\)-radius \(\ell_{\infty}\)-ball) on \(x\)

\[
\min_{\Theta} \mathbb{E}_{(x, y) \sim D}[\max_p L_{\Theta}(h(x + p), y)] \quad \text{where} \quad ||p||_{\infty} < \epsilon. \quad (1)
\]

From Eqn. 1, we observe that optimizing the robustness of neural networks is a saddle point problem. It can be decomposed as an inner maximization problem to strengthen the deception ability of the attack given the perturbation range \(\epsilon\), and an outer minimization problem to minimize the expected error given any attack.

B. Representative Attacks

FGSM [6] is a fast attack method that performs a one-step update using the sign of the gradient of the loss function, which increases loss in its sharpest direction.

\[
\hat{x} = x + \epsilon \cdot \text{sign}[\nabla_x L_{\Theta}(h(x), y)]. \quad (2)
\]

BIM [10] performs FGSM attack in multiple steps with a small step size \(\alpha\) and clips each perturbation update to a given range \(\epsilon\).

\[
\hat{x}_{t+1} = \text{Clip}\{\hat{x}_t + \alpha \cdot \text{sign}[\nabla_x L_{\Theta}(h(\hat{x}_t), y)]\}. \quad (3)
\]

PGD [15] is a variant of BIM that adds a random start. It projects the updated adversarial samples into an \(\epsilon\)-radius \(\ell_{\infty}\)-ball. PGD is proven to be a general first-order attack [15].

TPGD [29] is an improved version of PGD that considers robustness loss as the sum of the natural loss (classification loss) and the boundary loss through

\[
\hat{x}_t = \text{sign}[\nabla_x L_{\Theta}(h(\hat{x}_t), y) + L_{\Theta}(h(\hat{x}_t), h(x))/\rho], \quad (4a)
\]

\[
\hat{x}_{t+1} = \text{Clip}\{\hat{x}_t + \alpha \cdot \hat{x}_t\}, \quad (4b)
\]

where \(\rho\) models the tradeoff between the two loss terms. These four attack schemes are employed in our experiments.

On the defense side, adversarial training [15], [24] generates adversarial examples and then trains the model with them to improve robustness. Another branch of defense is randomization, which aims to randomize the adversarial perturbation into a random perturbation. Ref. [27] proposes a padding- and resize-based random input transformation to reduce adversarial perturbation. Random mask [14] filters the feature maps selectively with random masks, which helps a CNN capture spatial information and develop resistance to attacks.

C. Maximum Mean Discrepancy (MMD)

To bound the error in the target domain with assistance from source domains, MMD [17] has been extensively utilized as such a discrepancy metric.

Definition 2. Let \(\mathcal{H}\) be the Reproducing Kernel Hilbert Space (RKHS) endowed with mapping function \(\phi(\cdot) : X \rightarrow \mathcal{H}\). The MMD distance \(d_{MMD}(\cdot, \cdot)\) represents the distances between mean embeddings of feature generated by function \(\phi(\cdot)\). The expected MMD distance between two domains \(S\) and \(T\) is formalized by

\[
d_{MMD}(S, T) = \mathbb{E}_{(x, y) \sim S}[\phi(x) - \mathbb{E}_{(x, y) \sim T}[\phi(x)]]_{\mathcal{H}}, \quad (5)
\]

which usually induces kernel function \(k(x_1, x_2) = \langle\phi(x_1), \phi(x_2)\rangle_{\mathcal{H}}\) for MMD distance computation. In this work, the adopted kernel for \(\phi(\cdot)\) is the commonly used Gaussian kernel \(k(x_1, x_2) = \exp(-\frac{1}{2\sigma^2}||x_1 - x_2||^2)\) which is Lipschitz-continuous. MMD is employed to align the feature distribution between \(S\) and \(T\) in domain adaptation, which can be readily extended to domain generalization.

III. METHODOLOGY

Since the adversarially augmented data introduced during the robustness training have a different distribution from the natural data, and attacked data (from target domain) is unknown during training, it makes sense to view robustness training as domain generalization. To this end, we recast the goal from a domain generalization viewpoint. Here we denote each clean datum as \((x_0, y)\) drawn from a natural source domain \(S_0\), and the augmented adversarial samples during training are drawn from other source domains \(S_1, \ldots, S_{N_k}\).

Definition 3. Given \((N_k + 1)\) source domains, where natural data \((x_0, y) \sim S_0\), and multiple adversarial data \((x_{aadv}, y) \sim S_k\) for \(k = 1, \ldots, N_k\). The joint distribution \(P_{XY} \neq P_{XY}^j\) for any \(1 \leq i \neq j \leq N_k + 1\). The goal is to learn robust and quantized parameters \(\Theta\) that achieve minimum performance error on unseen attacked target data \((x_t, y) \sim T\), i.e. generated during inference.

\[
\min_{\Theta} \mathbb{E}_{(x_t, y) \sim T}[\max_p L_{\Theta}(h(x_t), y)], \quad (6)
\]

where the target input data \(x_t\) is generated by \(x_t = \arg\max_p L_{\Theta}(h(x_0 + p), y), ||p||_{\infty} < \epsilon\). Both model parameters \(\Theta\) and activations are limited to a small bitwidth (e.g., \(\leq 4\)-bit).
Our defensive framework

Representative defensive approaches

Fig. 2. Left: Conventional defensive approaches are time-consuming and spatially inefficient. The model is offline that requires far more iterations than natural training to generate adversarial samples and visibility of the whole dataset. Right: The proposed framework instantiates a global set that updates the inter-batch perturbation and utilizes the gradient information in one backward pass for weight update and local intra-batch perturbation generation in an online manner. The alignment loss (viz. MMD) is imposed on top to minimize the discrepancy of feature distributions between natural and adversarial data.

Algorithm 1 ODG-Q Training.

**Input:** Natural dataset \((X_0, Y)\) from domain \(S_0\).

**Parameter:** Number of online adversarial domains \(N_k\), local magnitude \(\epsilon_l\), perturbation bound \(\epsilon\), learning rate \(\eta\), trade-off factor \(\lambda\), quantized model \(h(\cdot)\) for brevity.

**Output:** A robust quantized model with parameter \(\Theta\).

1. Initialize global perturbation set \(P \in \mathbb{R}^{N_k \cdot B \cdot C \cdot H \cdot W}\) with zeros, \(k = 0\).
2. while epoch \(\leq N_e/2\) do
3. while batch \(\leq N_b\) do
4. \(p_k^b \leftarrow P[k, ...]\)
5. \(x_{gk} \leftarrow x_0 + \text{Clip}(p_k^b, -\epsilon, +\epsilon)\)
6. \(g_k^b \leftarrow \nabla_{\Theta} L_{task}(h(x_{gk}), y)\)
7. \(p_k^b \leftarrow \epsilon_l \cdot \text{sign}(g_k^b)\)
8. \(P[k : N_k - 1, ...] \leftarrow P[k : N_k - 1, ...] + p_k^b\)
9. \(x_{advk} \leftarrow x_{gk} + \text{Clip}(p_k^b, -\epsilon, +\epsilon)\)
10. \(g_{\Theta} \leftarrow \nabla_{\Theta} \left[ L_{task}(h(x_{advk}), y) + \lambda L_{MMD}(h(x_{advk}), h(x_0)) \right]\)
11. \(\Theta \leftarrow \Theta - \eta \cdot g_{\Theta}\)
12. end while
13. \(k \leftarrow (k + 1) \mod N_k\)
14. end while
15. return Robust quantized model with parameters \(\Theta\).

In the following, the loss function \(L_{\Theta}(\cdot, \cdot)\) is denoted as \(L\) for brevity. To improve the robustness of model against different attacks, a direct idea is to prepare diverse adversarial samples which have different times of adversarial attack iterations, different approaches of attacks, etc. Then adversarial samples can be combined with natural samples for training.

However, the aforementioned direct idea requires a huge space to store these adversarial samples, and expensive computation is needed for generating and training these adversarial samples. In addition, quantization/binarization introduces extra computation for all regular convolutional and fully-connected layers. To solve these problems, we propose that the adversarial data from multiple source domains \(S_1, ..., S_{N_k}\) are generated and updated sequentially on-the-fly during training, instead of preparing them before training.

As shown in Fig. 2, representative defensive approaches suffer from low time-space efficiency due to the multiple iterations offline to generate adversarial training samples. However, the visibility of the whole adversarial training dataset may not be necessary to the model. Inspired by online learning [8], the main drawback of offline approaches can be overcome if the adversarial training data stream can be generated online continuously.

Intuitively, the adversarial samples should meet two requirements: 1) they should be generated at a low computational cost, therefore extra training loops should be avoided; 2) they should be diverse within the \(\epsilon\)-radius \(\ell_\infty\)-ball of clean data. Therefore, the gradient ascent information is considered in both global-level (inter-batch) perturbation and local-level (intra-batch) perturbation, but not local-level perturbation only like in previous schemes [6], [15], [30].

To this end, we use a global perturbation set \(P \in \mathbb{R}^{N_k \cdot B \cdot C \cdot H \cdot W}\), where \(N_k\) is the number of online adversarial domains, and the remaining dimensions have the same shape as a batch of input data. The global perturbation set \(P\) accumulates the inter-batch gradient information. To make the adversarial samples diverse for producing different adversarial updates, we only update \(P[k : N_k - 1, ...]\) per epoch, where \(k\) is the epoch index modulo \(N_k\). By doing so, \(P[0, ...]\) updates every \(N_k\) epochs, and \(P[N_k - 1, ...]\) updates every epoch like a roulette. The adversarial source datasets \\{\((X_{advk}, Y) \sim S_k | k = 1, ..., N_k\)\} are thereby constructed by incorporating global perturbation \(P[k : N_k - 1, ...]\) and local perturbation on the natural dataset \((X_0, Y)\) sequentially. \(P\) helps to recycle the gradient information from perturbations generation to weight updating.

The proposed training process is summarized in Algorithm 1 dubbed “ODG-Q” wherein we efficiently generate diverse adversarial samples online. For each batch, in Stage 1 (lines 4–7), the global perturbation (inter-batch) in domain \(k\) is added to natural data to obtain \(x_{gk}\), then the local perturbation (intra-batch) is generated by gradient ascent on \(x_{gk}\). In Stage 2 (lines 8–9), the global perturbation set is updated with the
current local perturbation, and the adversarial data $x_{\text{adv}}$ is generated using global perturbation $p^g$ and local perturbation $p^l$. In Stage 3 (lines 10–11), the model is trained with $x_{\text{adv}}$. An MMD loss is added on the last convolution output for high-level feature alignment. Weight updates only once each batch without inner loop for adversarial samples generation. We halve the training epochs to keep our total training time consistent to natural training, since two backward passes are involved per batch in ODG-Q for perturbation generation and gradient descent, respectively.

The theoretical insight of ODG-Q, which discusses how the risk on attacked data during inference (data from target domain) can be bounded by the proposed algorithm, is described in Appendix due to space constraint.

IV. EXPERIMENTS

We conduct experiments on MNIST [11], CIFAR-10 [9] and ImageNet [4]. Quantization-aware training on uniform quantizer and XNOR-net [16] are employed for quantization and binarization (viz. 1-bit quantization), respectively. We quantize both weights and activation. We use different attacks for robustness evaluation, namely, FGSM [6], PGD [10], BIM [15] and TPGD [29]. “Natural” refers to the evaluation of the clean natural data. “GN” stands for Gaussian noise. PGD-20 means the number of perturbation updates is 20. The notation is similar for BIM-20 and TPGD-20, etc. ResNet-20 [7] is employed in MNIST and CIFAR-10. For ImageNet, we employ ResNet-18. Implementation details are given in Appendix.

A. Experimental Results

Table I reports the model accuracy under both white-box and black-box attacks, together with training time for different methods on CIFAR-10. The “Natural” in the first row means experiments conducted on the clean natural data, and “Natural” in the method column means natural training without any defense approach. From Table I, we observe that low-bitwidth neural networks are sensitive to adversarial attacks for CIFAR-10 dataset. Even Gaussian noise on the natural data causes performance degradation from 87.5% to 72.9%. K-iterative Training can improve robustness. However, the expensive computational cost makes it hard to scale to large datasets. AdvFree($m = 4$) and AdvFree($m = 6$) are two variants of AdvFree that set 4 and 6 times inner loops, respectively. The results of AdvFree show that it may not improve robustness effectively in the low-bitwidth setting. DQ is a regularization-based method having a small computational overhead, while its performance of robustness is not as good as training with the adversarial samples.

In contrast, our method 1) outperforms state-of-the-art methods in both accuracy and training time consistently by a large margin: achieving 49.2% average improvements under five white-box attacks and 21.7% under five black-box attacks in the setting of $N_k = 6$; 2) maintains good performance on...
TABLE III

| Defense Method | Natural | PGD-20 | BIM-20 | TPGD-20 | Imp.| PGD-20(B) | BIM-20(B) | TPGD-20(B) | Imp.| Total | Epoch |
|----------------|---------|--------|--------|---------|-----|----------|----------|----------|-----|-------|-------|
| Natural 4-bit  | 69.9    | 40.5/17.5 | 40.4/17.5 | 55.7/38.6 | 61.8/55.3 | 61.8/55.2 | 65.7/63.4 | 38.4 | 25.6 |
| Ours (Nk = 1)  | 69.9    | 45.6/21.2 | 43.6/21.4 | 59.3/42.1 | +3.4/+3.7 | 63.8/43.8 | 63.8/43.6 | 68.8/46.0 | +2.2/+3.1 | 42.7 | 36.9 |
| Ours (Nk = 2)  | 69.8    | 43.5/20.9 | 43.5/20.9 | 59.4/42.1 | +3.4/+3.5 | 63.8/45.8 | 63.9/45.0 | 68.3/45.8 | +2.1/+3.3 | 43.1 | 57.5 |
| Ours (Nk = 4)  | 69.9    | 43.8/21.3 | 43.7/21.3 | 59.3/42.4 | +3.4/+3.8 | 63.2/58.1 | 63.1/58.0 | 68.1/65.4 | +1.7/+2.5 | 42.7 | 56.9 |
| Ours (Nk = 6)  | 69.9    | 43.6/21.4 | 43.7/21.3 | 59.3/42.4 | +3.4/+3.9 | 64.4/60.0 | 64.3/60.0 | 68.6/66.3 | +2.7/+4.1 | 42.8 | 57.1 |

Natural T-bit

| Defense Method | Natural | PGD-20 | BIM-20 | TPGD-20 | Imp.| PGD-20(B) | BIM-20(B) | TPGD-20(B) | Imp.| Total | Epoch |
|----------------|---------|--------|--------|---------|-----|----------|----------|----------|-----|-------|-------|
| Ours (Nk = 1)  | 47.3    | 44.0/41.2 | 44.0/41.2 | 48.5/44.0 | +2.4/+3.3 | 36.7/28.8 | 36.7/28.9 | 44.3/40.2 | +2.3/+2.5 | 40.2 | 53.6 |
| Ours (Nk = 2)  | 47.2    | 44.5/41.9 | 44.7/41.9 | 46.7/44.9 | +2.9/+4.0 | 36.0/28.5 | 36.0/28.5 | 43.5/39.2 | +1.6/+1.7 | 40.3 | 53.8 |
| Ours (Nk = 4)  | 47.3    | 44.2/41.9 | 44.3/41.8 | 46.0/44.5 | +2.4/+3.9 | 36.1/27.7 | 36.1/27.7 | 43.9/39.6 | +1.9/+1.3 | 39.9 | 53.3 |
| Ours (Nk = 6)  | 47.3    | 44.2/41.6 | 44.3/41.4 | 46.7/44.3 | +2.7/+3.6 | 36.9/28.9 | 36.9/28.9 | 44.1/40.2 | +2.4/+2.3 | 40.0 | 53.3 |

Ablation study for different variants of our model. Top-1(%) accuracy of 4-bit ResNet-18 on CIFAR-10 against various adversarial attacks ($\epsilon = 8$). The mark “global pert.” means ODG-Q removes the inter-batch global perturbation, and “MMD (0.3)” means we replace the coefficient of MMD loss $\lambda$ from 3.0 to 0.3.

Method | Natural | GN | FGSM | PGD-20 | BIM-20 | TPGD-20 | Imp.| GN(B) | FGSM(B) | PGD-20(B) | BIM-20(B) | TPGD-20(B) | Imp.(B) |
|--------|---------|----|------|--------|--------|---------|-----|-------|--------|---------|--------|---------|-------|
| Ours (Nk = 4) | 84.9 | 82.2 | 81.3 | 81.1 | 81.2 | 82.4 | 81.6 | 46.7 | 46.8 | 61.1 |
| Ours w/o global pert. | 71.1 | 75.1 | 77.1 | 77.1 | 76.9 | 51.1 | 40.8 | 63.9 | -4.2 |
| Ours w/o MMD | 88.7 | 76.6 | 40.2 | 18.8 | 18.9 | 49.2 | -41.1 | 76.6 | 53.6 | 39.8 | 39.9 | 62.5 | -6.4 |
| Ours w/ MMD (0.3) | 86.5 | 81.7 | 66.1 | 59.2 | 59.2 | 73.7 | -25.8 | 81.7 | 53.7 | 36.5 | 36.8 | 60.6 | -7.1 |

TABLE IV

Visualizing what the model learns in natural domain $S_0$ and online adversarial domains $S_{adv}$ and $S_{adv1}$. The true label of example is tench (a kind of fish). 1) All the natural and adversarial data focus on the area containing the region of interest (RoI) successfully. 2) Fewer high-confidence areas (warm color) focus on the areas without RoI in adversarial domains than that in the natural domain $S_0$. Right: T-SNE visualization [25] of features for natural images (red), FGSM-attacked images (blue), PGD-attacked images (green) and TPGD-attacked images (purple). Regardless of bitwidths, the features have large discrepancies in natural training, while being aligned well in ours, which shows why our method enables robust performance on the attacked images.

![Image](image.png)

(a) $N_k = 2$ Multi-domain Online Samples Visualization

(b) T-SNE visualization of feature distributions

The performance on MNIST is reported in Table II. Our method achieves good accuracy under various attacks, with efficient training time per epoch. In Table III, we report the defense performance on the large-scale ImageNet dataset in both 4-bit and 1-bit settings. Due to the high computational requirement, few teams have attempted the adversarial robustness on the ImageNet dataset, let alone training in the low-bitwidth setting which also requires training extra quantization layers. To our best knowledge, this paper is the first to train both quantized and binary neural networks on ImageNet successfully against adversarial attacks. As shown in Table III, our method boosts the robustness to PGD, BIM and TPGD attacks steadily, with the total training time close to that of natural training. Since both the adopted quantization/binarization method and the model architecture (ResNet-18) in this paper are basic, there exists room for improvement for the ImageNet dataset.
TABLE V

PERFORMANCE COMPARISON BETWEEN NATURAL TRAINING AND OURS
(Nk = 4) WITH DIFFERENT BITWIDTHS ON CIFAR-10 DATASET.

| Bit width | Method       | Natural | FGSM  | PGGD-20 | BIM-20 | TPGGD-20 | Imp ↑ |
|----------|--------------|---------|-------|---------|--------|----------|------|
|          | Natural      | 88.0    | 76.1  | 42.3    | 21.2   | 21.2     | 49.8 |
| 32bit    | Ours         | 88.0    | 82.1  | 82.5    | 82.5   | 83.8     | -    |
| 8bit     | Ours         | 86.9    | 77.5  | 81.5    | 81.5   | 81.5     | +40.5|
| 4bit     | Ours         | 85.6    | 79.7  | 84.4    | 84.4   | 84.4     | +42.0|
| 3bit     | Ours         | 84.5    | 80.2  | 85.1    | 85.1   | 85.1     | +49.2|
| 2bit     | Ours         | 82.7    | 79.6  | 85.4    | 85.4   | 85.4     | +53.5|
| 1bit     | Ours         | 81.2    | 64.8  | 55.1    | 55.1   | 55.1     | +26.3|

A. Ablation Study of ODG-Q

We provide an ablation study of ODG-Q in Table IV. In the setting “w/o global pert.”, we remove the global perturbation in Algorithm 1. The training is still in an online version that adversarial samples are generated dynamically during training. However, only local perturbation generated from the data itself can be used for adversarial training, which is also adopted in previous methods [20], [6], [15], [30]. The performance gap between “w/o global pert.” and the full ODG-Q demonstrates the importance of global perturbation from different batch data, which gives the model a holistic view of data and helps the model to generate diverse adversarial samples in different source domains, i.e. the domain of source data depends on the times of using and updating global perturbation set during training. The visualization of different source data is shown in Fig. 3(a). The visualization shows that the data from different source domains have different gradient information.

In the remaining settings, we remove the MMD loss or set a small MMD loss coefficient. We can observe that MMD loss plays a significant role during training. Without the feature alignment between natural samples and various online adversarial samples by MMD loss, the training may become unconstrained.

B. Performance in Different Bitwidths

We further test the performance of our method when quantizing the network to different bitwidths. Although this paper focuses on robustness improvement for quantized and binary models, the proposed Algorithm 1 is orthogonal to the used bitwidth. Therefore, we also experiment on full-precision model (32-bit), besides low-precision settings ranging from 8-bit down to 1-bit. From Table V, our approach promotes robustness against different adversarial attacks by a large margin. The robustness improves significantly not only at full-precision settings, but more importantly, in quantized and binarized settings. We also observe that the level of improvement increases from 32-bit to 4-bit, and decreases from 4-bit to 1-bit. It indicates that the robustness improvement becomes a little easier when quantized to 8-bit, demonstrating that small adversarial perturbations can be absorbed in this case. However, in a low-bitwidth setting (≤ 4-bit), robustness learning becomes very challenging with a tightened bitwidth. It is reasonable since the capability of a model is strictly constrained by the low-bitwidth setting such that even training the model with diverse adversarial samples, the performance is limited.

C. Interpretablity of Proposed Robust Training

In Fig. 3(a), we visualize what the model has learned from the source domain $S_0$ and online adversarial domains $S_{adv1}$ and $S_{adv2}$. With small perturbations on the natural images, Grad-CAM [18] and Backpropagation Saliency map [21] are adopted to check the attention captured by the model. Interestingly, we find that there are fewer high-confidence areas that focus on the areas without the region of interest (RoI) in adversarial domains than that in the natural domain $S_0$. For example, the Grad-CAM in $S_0$ treats not only the fishes themselves but also surrounding water areas as high-confidence areas, while the Grad-CAM in $S_{adv1}$ and $S_{adv2}$ tends to have lower interest in the surrounding water area. It hints that training with the online adversarial samples improves robustness by encouraging the model to focus more on the useful RoI.

We visualize the T-SNE projected features distribution of natural training, 5-PGD training and the proposed ODG-Q on CIFAR-10 dataset in Fig. 3(b). It is observed that the features of natural images (red) is quite different from the features of various attacked images, thereby the predictor using natural training can be deceived to give incorrect predictions on attacked images. Instead, by training natural domains and online adversarial domains altogether with MMD loss, the high-level features from different domains are aligned closer together, thus allowing the proposed model to give correct predictions regardless of the attacks.

VI. CONCLUSIONS

In this paper, we novelly recast robustness training as a domain generalization problem in an online version where multiple source data are generated dynamically. We then propose an strategy to generate diverse source data considering both global and local perturbations. Powered by the MMD loss, the training process is safely constrained that high-level features of the natural and adversarial online data are aligned, which reduces the expected risk of model on attacked data during inference in theory. Extensive experiments show that the proposed ODG-Q defeats various attacks with a big margin over existing defense schemes in both white-box and black-box settings, using only minimal computation and a training time near natural training. We are hopeful that this work bridges the gap between network quantization and adversarial robustness.

ACKNOWLEDGEMENTS

This work is supported in part by the General Research Fund (GRF) project 17206020, and in part by ACCESS, AI Chip Center for Emerging Smart Systems, Hong Kong SAR.
