Feature Distillation: DNN-Oriented JPEG Compression Against Adversarial Examples

Zihao Liu[1], Qi Liu[1], Tao Liu[1], Nuo Xu[1], Xue Lin[2], Yanzhi Wang[2], Wujie Wen[1]
[1] Flordia International University, [2] Northeastern University
{zliu021,qliu020,tliu023,nxu003,wwen}@fiu.edu, {xue.lin, yanz.wang}@northeastern.edu

Abstract

Image compression-based approaches for defending against the adversarial-example attacks, which threaten the safety use of deep neural networks (DNN), have been investigated recently. However, prior works mainly rely on directly tuning parameters like compression rate, to blindly reduce image features, thereby lacking guarantee on both defense efficiency (i.e. accuracy of polluted images) and classification accuracy of benign images, after applying defense methods. To overcome these limitations, we propose a JPEG-based defensive compression framework, namely "feature distillation", to effectively rectify adversarial examples without impacting classification accuracy on benign data. Our framework significantly escalates the defense efficiency with marginal accuracy reduction using a two-step method: First, we maximize malicious features filtering of adversarial input perturbations by developing defensive quantization in frequency domain of JPEG compression or decompression, guided by a semi-analytical method; Second, we suppress the distortions of benign features to restore classification accuracy through a DNN-oriented quantization refine process. Our experimental results show that proposed "feature distillation" can significantly surpass the latest input-transformation based mitigations such as Quilting and TV Minimization in three aspects, including defense efficiency (improve classification accuracy from $\sim 20\%$ to $\sim 90\%$ on adversarial examples), accuracy of benign images after defense ($\leq 1\%$ accuracy degradation), and processing time per image ($\sim 259\times$ Speedup). Moreover, our solution can also provide the best defense efficiency ($\sim 60\%$ accuracy) against the recent adaptive attack with least accuracy reduction ($\sim 1\%$) on benign images when compared with other input-transformation based defense methods.

1. Introduction

Recent studies have shown that DNN models are inherently vulnerable to adversarial examples (AEs) [10, 24], i.e. malicious inputs crafted by adding small and human-imperceptible perturbations to normal inputs, strongly fooling the cognitive function of DNNs. For example, in image recognition, adversarially manipulating the perceptual systems of autonomous vehicles by physically captured adversarial images, i.e. via camera or sensor [17, 22], can lead to the misreading on road signs, thus causing potential disastrous consequences in DNN-based cyber-physical systems.

Many countermeasures [15, 14, 20, 23, 25, 2] have been proposed to enhance the robustness of DNNs against adversarial examples, mainly including DNN model-specific hardening strategies and model-agnostic defenses [11]. Typical model-specific solutions like "adversarial training" or "defensive distillation" may rectify the model parameters to mitigate the attacks by using the iterative retraining procedure or masking adversarial gradient. The model-agnostic approaches such as input dimensionality reduction [5, 26] or direct JPEG compression [9, 7, 11] attempt to remove adversarial perturbations from the inputs, before feeding them into DNN classifiers.

In this work, we focus on improving the effectiveness and efficiency of compression based model-agnostic mitigation against adversarial examples. Though standard JPEG compression has been explored to mitigate the adversarial examples [9, 7], it can neither effectively remove the adversarial perturbations, nor guarantee the classification accuracy on benign images, due to its focus on human visual quality. Instead, we propose the DNN-favorable JPEG compression, namely *"feature distillation"*, by redesigning the standard JPEG compression algorithm, in order to maximize the defense efficiency while assuring the DNN testing accuracy. In specific, 1) We reveal the root reason to limit the JPEG defense efficiency by analyzing the frequency feature distributions of adversarial input perturbations during JPEG compression; 2) Inspired by our observation, we propose a semi-analytical method to guide the defensive quantization process to maximize the effectiveness of filtering adversarial features; 3) We characterize the importance of input features for DNNs by leveraging the statistical frequency component analysis within JPEG, and then develop DNN-oriented quantization method to recover the degraded

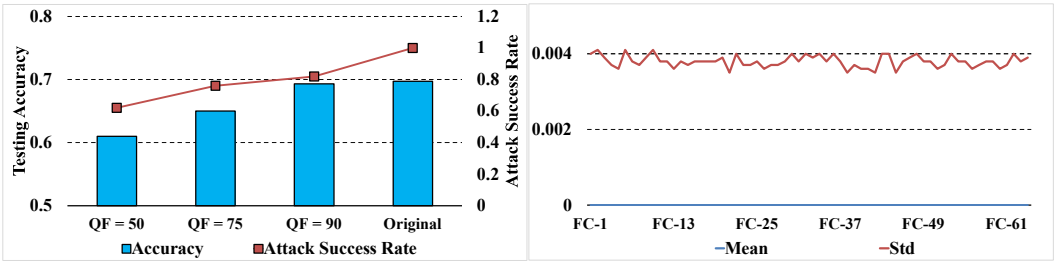

Figure 1: (a) Testing accuracy v.s. attack success rate at different QFs of JPEG; (b) Statistical information of FGSM-based AE perturbations in frequency domain (FC denotes frequency component)

accuracy (i.e., a side-effect induced by the feature loss in perturbation removal) on benign samples.

Our proposed method is built upon the light modifications of widely adopted JPEG compression and does not require any expensive model retraining or multiple model predictions. Evaluations show that *"feature distillation"* offers significantly improved effectiveness against a variety of mainstream adversarial examples (i.e., $> 90\%$ accuracy on AEs), with very marginal accuracy reduction (i.e., $\leq 1\%$) on benign data. Besides, it well beats recent proposed image transformation based defense like Quilting and TV Minimization in terms of defense efficiency, accuracy and processing speed. Furthermore, our solution offers the best defense efficiency ($\sim 60\%$) with lowest accuracy loss ($\leq 1\%$) against the recent adaptive attack–Backward Pass Differentiable Approximation (BPDA) [3] among existing input-transformation based defenses, though it is not completely immune to such attack. *To our best knowledge, there is no published work that can completely mitigate BPDA, since it is very challenging for defense if attackers can iteratively strengthen the adversarial examples according to the defense. However, we believe our work provides a new angle to redesign input-based defense to well balance the accuracy of benign data and defense efficiency with DNN-oriented/defensive quantization. It is a new trial towards developing better input-transformation based defenses.*

2. Background, Related Work and Motivation

2.1. Basics of Adversarial Examples and JPEG

Adversarial examples: ($X^* = X + \delta_X$) are created to fool the DNNs ($Y^* \neq Y$) with imperceptible perturbations: $\arg\min_{\delta_X} \parallel \delta_X \parallel \ s.t.\ F^{(\Theta)}(X + \delta_X) = Y^*$, which can be solved through many crafting algorithms: 1) **FGSM** [10] (fast gradient sign method) is a L_∞ attack and utilizes the gradient of the loss function to determine the direction to modify all the input pixels. It is designed to be fast, rather than optimal; 2) **BIM** [13] (basic iterative gradient sign method) is the iterative version of FGSM by gradually adding small perturbations α (L_∞) until reaching the upper bound ϵ or achieving successful attack; 3) **Deepfool** [16] uses geometrical knowledge to search the minimal perturbations (L_2) by assuming DNN as a linear classifier and each class is separated by a hyper-plane. Such an approach finds the nearest hyper-plane from X and uses geometrical knowledge to calculate the projection distance; 4) **C&W** [6] (Carlini & Wagner method) are a series of L_0, L_2, and L_∞ attacks that achieve 100% attack success rate with much lower distortions comparing with the above-mentioned attacks. In particular, the C&W L_2 attack uses a more effective objective function $f(x) = max\,(max\,\{Z(X)_i \ \ i \neq t\} - Z(X)_t, -\kappa)$ with logits $Z(X)_i$ and adjustable parameter κ. Further, C&W L_0 and L_∞ attacks are implemented indirectly by iteratively calling their L_2 attack. 5) **BPDA** [3] is the latest adaptive attack by recurrently computing the adversarial gradient after applying defense: $x^* = clip(x + \epsilon \cdot sgn(\nabla_x J_{\theta,Y}(DEF(x))))$, where J represents the function of an DNN model and DEF is the applied defense method in BPDA attack. It is state-of-the-art of attack by assuming adversaries know the defense method.

JPEG: [27] is a popular lossy compression standard for digital images based on the fact that Human-Visual System (HVS) is less sensitive to the high frequency components than low frequency ones [31]. A typical JPEG compression mainly consists of image partitioning, discrete cosine transformation (DCT), quantization, zig-zag reordering and entropy coding, etc. [27]. To compress a raw image, the high (low) frequency DCT coefficients are usually scaled more (less) and then rounded to nearest integers by performing element-wise division based on a predefined 8×8 Quantization Table (Q-Table) [27]. The trade-off between image quality and compression rate is realized by scaling each element in Q-Table via the "Quantization Factor" (QF) [30]. A higher compression rate corresponds to a lower QF. A reverse procedure of above steps can decompress an image.

2.2. Related Works

Applying JPEG compression to mitigate adversarial examples has been discussed in prior work. Kurakin et al. [13] test some model-agnostic approaches on adversarial examples and reveal a good potential of JPEG compression

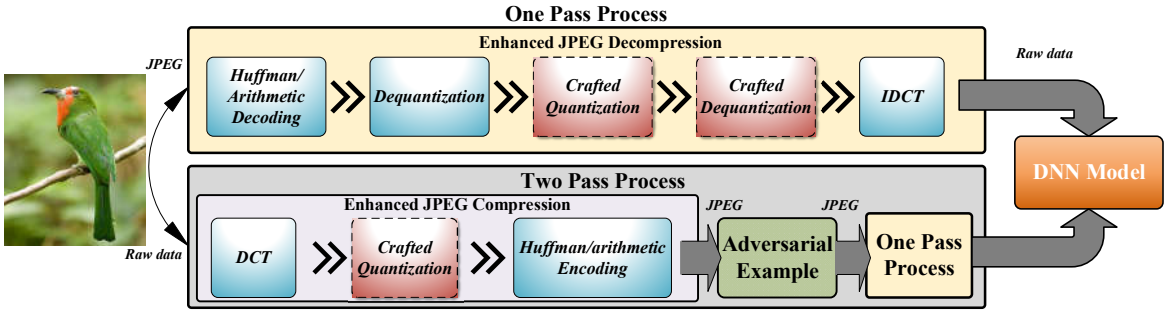

Figure 2: Illustration of two different modes of "feature distillation"–one pass and two pass.

for defending adversarial attacks. Dziugaite et al. [9] empirically report JPEG compression can reverse only small adversarial perturbations, but the reason behind is uncertain. Guo et al. [11] test JPEG compression, image Quilting (piecing together small patches from a database of image patches), total variance minimization (combining pixel dropout with total variation minimization), etc. against the gray-box and black-box adversarial attacks, and report Quilting and TVM show better efficiency than JPEG. Aydemir et al. [4] compare the effects of JPEG compression and JPEG2000, against adversarial perturbations. Though JPEG2000 shows better performance than JPEG, the efficiency is still far from satisfactory. Xu et al. [29] propose an ensemble method, namely "feature squeezing", to defeat the adversarial examples by integrating different types of "squeezers" (i.e. model-agnostic processing). Das et al. [7] propose a JPEG compression based ensemble method, namely "vaccinating", to mitigate adversarial attacks by voting the result based on a variety of compression rates. Prakash et al. [18] develop "pixel deflection" and "adaptive soft-thresholding" approaches by replacing or smoothing adversary perturbations. This method shows good defense efficiency on gray box-setting without evaluating adaptive attacks. Xie. et al. [28] propose two randomization operations–random size and random padding, against adversary examples. *In summary, prior studies empirically test the JPEG compression by directly tuning the compression rate, without digging into the underlying image processing mechanisms. The conclusion is that JPEG suffers from very limited defense efficiency but inevitable DNN accuracy degradation. To overcome those issues, standard JPEG compression should be integrated with the costly ensemble solutions. On the other side, our work directly targets the fundamental entities of JPEG compression/decompression, like defensive and DNN-oriented quantizations, to unleash its defense potentials with almost zero loss of DNN testing accuracy, thus is low-cost.*

2.3. Why standard JPEG is not good?

DNN suffers from both low testing accuracy and weak defense efficiency against adversarial examples if we directly employ standard JPEG compression based on human-visual system (HVS). To explore how existing compression can impact DNN's testing accuracy, we trained a MobileNet [12] with high quality JPEG images (QF=100, ImageNet), and tested it with both clear images and FGSM-based adversarial examples at various QFs (i.e., QF=100, 90, 75, 50). As Fig. 1 (a) shows, the testing accuracy degrades significantly as the compression rate increases (or QF from 100 to 50), despite the slightly improved defense efficiency (or drop in attack success rate). To achieve the best defense efficiency among our selected four QFs (attack success rate = 0.62 at QF = 50), the accuracy is even reduced by $\sim 8\%$ on benign images than that of the original one (QF=100). Apparently, the HVS-based JPEG compression is not an ideal solution in terms of defense efficiency and accuracy. Fig. 1 (b) further shows the means and standard deviations of DCT coefficients of malicious distortions at all 64 frequency bands. Given that malicious perturbations are almost randomly distributed in every frequency band, HVS-based JPEG compression, which distorts more (less) on high (low) frequency components of the input, is unlikely to effectively filter the distortions across the whole spectral domain.

3. Our Approach–Feature Distillation

In this section, we first provide a detailed analysis on how to wisely redesign the quantization process in JPEG compression to minimize attack success rate. As this lossy compression will still reduce the classification accuracy (see Fig. 1), we then develop the DNN-oriented quantization refine method, to compensate the reduced accuracy of benign images. Based on how/where the derived quantization will be placed in JPEG, our framework supports two modes (see Fig. 2): 1) **One pass process** by inserting a new quantization/de-quantization only in the decompression of standard JPEG; 2) **Two pass process** by also replacing the quantization of compression, followed by one pass process. The two pass method provides an opportunity to directly embed crafted quantization at sensor side to compress raw data to further improve defense efficiency,

given that JPEG-based image compression, an integrated component in sensors, is usually a "must-have" step to save data storage/transfer cost in real applications. *Therefore, the one-pass handles incoming images compressed by standard JPEG before sending them to DNNs, while the two-pass targets raw data directly sampled by devices like image sensors.* The target is to address both attack efficiency and test accuracy simultaneously.

3.1. Step 1: Defensive Quantization for Enhancing Defense Efficiency

We propose to use spectral filter by leveraging quantization process in JPEG on DNN inputs (i.e., adversarial examples), in order to mitigate adversarial perturbations.

One pass process. The idea is to directly filter out the malicious perturbations in frequency domain through the quantization process. As Fig. 2 shows, the JPEG-formatted input will be decompressed and then feed into the DNNs as the raw data at the beginning. By taking this chance, we insert a new pair of quantization/dequantization processes after the dequantization of standard JPEG decompression to purify the potential adversarial perturbations. Note we omit the first dequantization in the following analysis ideally by assuming it can almost preserve all frequency features of the input. Assuming for each 8×8 block in the input image X, adversarial distortion δ_X is added to X with intensity ϵ. The DCT transformation–a linear operation, essentially projects the image from spatial domain to spectral domain. Therefore, the original input and adversarial perturbations could be linearly separated as:

$$DCT(X+\delta_X) = DCT(X) + DCT(\delta_X) = C_X \cdot B + C_{\delta_X} \cdot B \quad (1)$$

where C_X and C_{δ_X} are the DCT coefficients of X and δ_X, respectively, for the 8×8 image block, and B is the DCT transformation basis. The maximum magnitude of C_{δ_X} can be calculated by the summation of all 64 frequency components and each term is bounded by $cos(\theta) \cdot \epsilon$. Thus we have $-8 \cdot \epsilon < C_{\delta_X} < 8 \cdot \epsilon$. The DCT coefficients will be quantized again in this decompression process, providing a good opportunity for filtering the perturbations. The quantization is approximated as:

$$Round\left(C_X + C_{\delta_X}/QS\right) \approx Round\left(C_X/QS\right) + Round\left(C_{\delta_X}/QS\right) \quad (2)$$

where QS is the defensive quantization step (QS). Ideally, if $QS > |C_{\delta_X}|$, then the perturbation C_{δ_X} can be eliminated. However, this equation may induce undesired rounding error to limit the efficiency of removing malicious perturbations, given that C_{δ_X} is usually much smaller than C_X. We further analyze such a rounding error by decomposing $C_X = \eta + {}^{QS}/_2$, then we have:

$$Round\left(C_X + C_{\delta_X}/QS\right) = Round\left(\eta + {}^{QS}/_2 + C_{\delta_X}/QS\right) \quad (3)$$

If ${}^{QS}/_2 + C_{\delta_X} > QS$, this part will be rounded to $\pm 1, \pm 2, \pm 3...$, which will induce a stronger rounding error than the adversarial perturbations, resulting in degraded defense efficiency.

Two pass process. To avoid such rounding error, we further propose two pass method. As Fig. 2 shows, the raw data (i.e. sampled by sensors) will be compressed through a defensive quantization process, rather than the standard JPEG quantization, followed by an entire one pass process. Assuming such compressed benign inputs are then polluted by adversarial perturbations, adversarial examples will be further processed by considering both compression/decompression procedures as:

$$Round\left(\left(Round\left(\frac{\eta + QS/2}{QS}\right) * QS + C_{\delta_X}\right)/QS\right) = Round(\eta) \quad (4)$$

The malicious perturbations can be appropriately filtered without inducing any rounding error if QS satisfies the following equation:

$$Round\left(C_{\delta_X}/QS\right) = 0 \Rightarrow QS > 2\left|C_{\delta_X}\right|, C_{\delta_X} \in \left(-8\epsilon < C_{\delta_X} < 8\epsilon\right) \quad (5)$$

Therefore, we adopt the same QS ($QS > 16 \cdot \epsilon$) to eliminate the perturbations C_{δ_X} in both passes.

3.2. Step 2: DNN-Oriented Quantization for Compensating Accuracy Reduction

To recover the testing accuracy (see Section 2.3), our next step is to develop a DNN-oriented JPEG compression method by refining the defensive quantization table from step 1. We analyze the difference between human visual system (HVS) and DNN on feature extractions, and then propose a heuristic design flow.

Difference between HVS&DNN on Feature Extractions. Since the feature loss happens in the frequency domain after the DCT process, we first study the problem that which frequency components have the most significant impact on DNN results. Assume x_k is a single pixel of a raw image X, and x_k can be represented by 8×8 DCT:

$$x_k = \sum_{i=0}^{7} \sum_{j=0}^{7} c_{(k,i,j)} \cdot b_{(i,j)} \quad (6)$$

where $c_{(k,i,j)}$ and $b_{(i,j)}$ are the DCT coefficient and its basis function at 64 different frequencies, respectively. It is well known that the human visual system (HVS) is less sensitive to high frequency components but more sensitive to low frequency ones. The JPEG quantization table is designed based on this fundamental understanding. However, DNNs examine the importance of the frequency information in a quite different way. The gradient of the DNN function F with respect to a basis function $b_{(i,j)}$ is calculated as:

$${}^{\partial F}/_{\partial b_{(i,j)}} = {}^{\partial F}/_{\partial x_k} \times {}^{\partial x_k}/_{\partial b_{i,j}} = {}^{\partial F}/_{\partial x_k} \times c_{(k,i,j)} \quad (7)$$

Eq. (7) implies that the contribution of a frequency component ($b_{i,j}$) to the DNN result will be mainly decided by

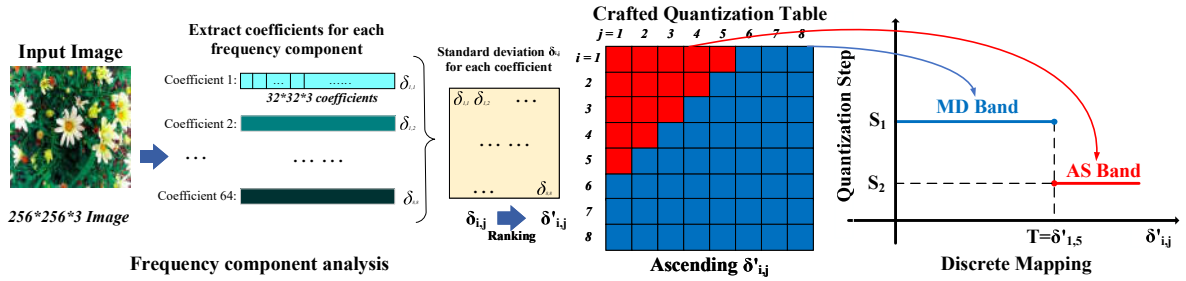

Figure 3: An overview of heuristic design flow of DNN-Oriented compression based on crafted quantization.

its associated DCT coefficient ($c_{(k,i,j)}$) and the importance of the pixel ($\partial F/\partial x_k$). Here $c_{(k,i,j)}$ will be distorted by the quantization before training. Ideally a well trained DNN model should respond with different strengths to all the 64 frequency components depending on the $c_{(k,i,j)}$ values. From this observation, large $c_{(k,i,j)}$ should be compressed less (using a small quantization step) in order to ensure a desirable classification accuracy.

In contrast, the default quantization table used in JPEG focuses on compressing more on less sensitive frequency components to HVS. As a result, in order to defend against adversarial attacks, aggressive compression is required, making DNNs easily misclassified if original versions contain important high frequency features. The DNN models trained with original images learn comprehensive features, especially high frequency ones. However, such features are actually lost in more compressed testing images, resulting in considerable misclassification rate (see Fig. 1(a)).

Therefore, we propose to compensate the accuracy reduction incurred by defending adversarial examples through a heuristic design flow (see Fig. 3): 1) characterize the importance of each frequency component through frequency analysis on benign images; 2) lower the quantization step of the most sensitive frequency components based on the statistical information for enhancing accuracy.

A: Frequency Component Analysis. For each input image, we first characterize the pre-quantized DCT coefficient distribution at each frequency component. Such a distribution represents the energy contribution of each frequency band [19]. Prior work [19] has proved that the pre-quantized coefficients can be approximated as normal (or Laplace) distribution with zero mean but different standard deviations ($\delta_{i,j}$). A larger $\delta_{i,j}$ means more energy in band (i,j), hence more important features for DNN learning. As Fig. 3 shows, each image will be first partitioned into N 8×8 blocks, followed by a block-wise DCT. Then the DCT coefficient distribution at each frequency component will be characterized by sorting all coefficients at the same frequency component across all image blocks. The statistical information, such as the standard deviation $\delta_{i,j}$ of each coefficient, will be calibrated from each individual histogram.

B: Quantization Table Refinement. Once the importance of frequency components is identified based on the standard deviations of DCT coefficients ($\delta_{i,j}$), the next step is to boost the accuracy of legitimate examples $\{acc_l\}$ (refer to the testing accuracy of benign images processed after the defense method). Our analysis in Section 2.3 indicates that a proper selection of QS can effectively mitigate the perturbations, whereas larger QS will induce more quantization errors. Therefore, we reduce the quantization errors of the most sensitive frequency components to enhance the testing accuracy by lowering their corresponding quantization steps within the quantization table, but such frequency components should be as few as possible to maintain the defense efficiency. In specific, we first sort the magnitude of $\delta_{i,j}$ in an ascending order as $\delta'_{i,j}$, then set the appropriate quantization step based on $\delta'_{i,j}$. To simplify our design, we introduce a discrete mapping function to derive the quantization step on each frequency band, base on the associated standard deviation $\delta_{i,j}$, i.e., $QS_{i,j} = (\delta_{i,j} \leq T \ ? \ S_1 : S_2)$, where T is the threshold to divide the 64 frequency components. Note that $S_1 > S_2$. The 64 frequency components are divided into two bands (see Fig. 3): the red colored Accuracy Sensitive (AS) band with $QS = S_2$, and the blue colored Malicious Defense (MD) band with $QS = S_1$ from Section 3.1.

4. Evaluation

In this section, we first explore the parameter optimization in our feature distillation under the constraints of high classification accuracy on malicious inputs after applying defense, while preserving the accuracy of legitimate ones given that both types of data can arrive for a realistic DNN testing. Then we comprehensively evaluate feature distillation under following three different settings: 1). **Gray-box:** We assume the adversary has full access to DNN model, but is unaware of the input transformations applied (defense method unaware) [11, 8]. 2). **White-box:** We consider adversary has full access to the DNN model, as well as the full knowledge of the defense method [3], which is more challenging. 3). **Black-box:** We assume adversary does not know the exact network architecture and weights, instead, can use a substitute model to craft adversarial perturbations that are transferable to the target model [11].

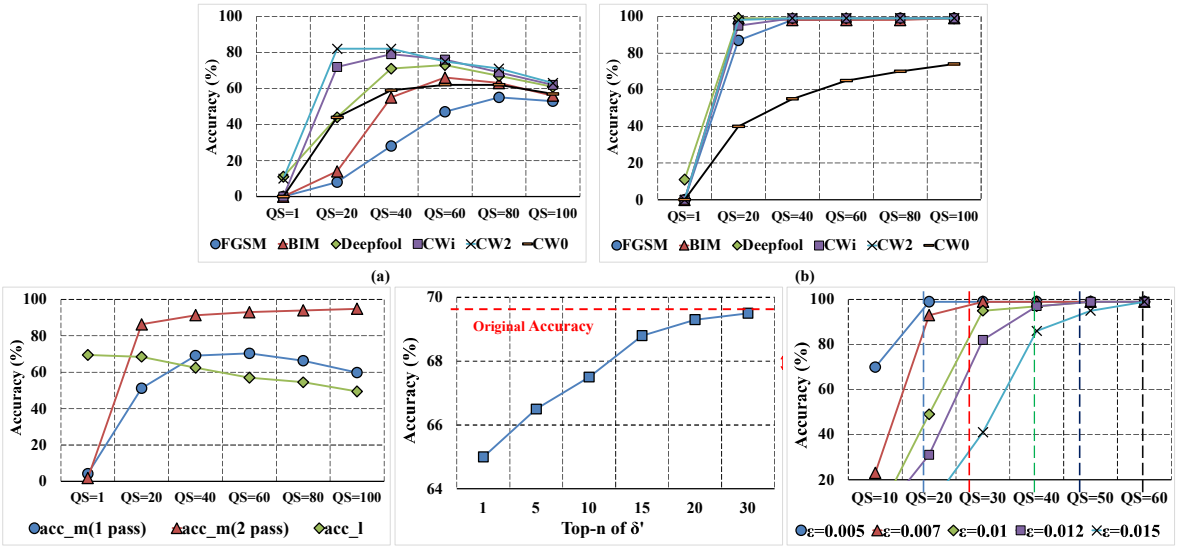

Figure 4: Exploration of the defensive quantization step for a 8*8 table: (a) Defense efficiency of one pass method against adversarial examples; (b) Defense efficiency of two pass method against adversarial examples; (c) Average defense efficiency w.r.t. the legitimate image accuracy (FGSM, $\epsilon = 0.008$); (d) Accuracy impacts of ranked frequency components (FGSM, $\epsilon = 0.008$); (e) Accuracy impacts of various quantization steps w.r.t. different perturbation strength (FGSM).

4.1. Experimental Setup

Our experiments are conducted on the Tensorflow DNN computing framework [1], running with Intel(R) Xeon(R) 3.5GHz CPU and two parallel GeForce GTX 1080Ti GPUs. Our proposed feature distillation method is implemented on the heavily modified adversarial machine learning library–EvadeML-Zoo [29] for white and gray-box settings and BPDA attack [3] for white box setting. To better illustrate the image compression based mitigation, we choose the large-scaled ImageNet dataset as our benchmark. Four other input-based countermeasures, including default JPEG [9, 13], bit-depth (one of the feature squeezing methods by reducing the bit number of an image pixel) [29] and the recent proposed TV Minimization (TVM) and Image Quilting [11] , are selected as the baselines to compare with our proposed feature distillation.

Methodology. Various types of adversarial example attacks, i.e., FGSM, BIM, Deepfool, CW_2, CW_0, CW_∞ and adaptive attack–BPDA, have been simulated in our experiments for evaluating the defense. We adopt a similar evaluation model from [29]. First, we choose 1000 benign images (one per class) to evaluate the testing accuracy of each DNN model. The seed images, which will be adding adversarial perturbations, are selected from the first 100 correctly predicted examples in the 1000 selected images on each DNN model for all the attack methods. The legitimate examples *classification accuracy* (acc_l) is the testing accuracy of benign images processed by the defense method. The *defense efficiency* is measured by the classification accuracy (acc_m) of 100 polluted images after applying the defense method.

4.2. Optimized Quantization Step

Defending against adversarial examples. Fig. 4 (a) and (b) illustrate the impact of the quantization steps of the 8*8 table under various adversarial attacks with our one-pass and two-pass defense approaches applied, respectively. Apparently, both processes demonstrate that the defense efficiency can be steadily improved as the QS grows, however, it will be saturated (even decreased) if QS becomes too large for the two pass (one pass) process. Compared with the one pass process, the two pass process always delivers much better defense efficiency against most of the adversarial attacks (except the CW_0), due to the elimination of the rounding error. The reason is because CW_0 attacks attempt to use a minimum number of pixel(s) with maximum perturbations to fool the DNN models, therefore the perturbations of each single pixel will translate into larger magnitudes than the other attacks in the frequency domain. This leads to a much higher QS for completely removing the associated perturbations, as Fig. 4 (a) and (b) show.

Evaluating testing accuracy. Fig. 4 (c) shows the testing accuracy changes w.r.t. QSs for both malicious examples (acc_m) and legitimate examples (acc_l). The acc_m (1 pass) and acc_m (2 pass) represent the average accuracy of various adversarial examples by applying our one pass and two pass process, respectively. As Fig. 4 (c) shows, acc_m (1 pass) and acc_l demonstrate an opposite trend as QS grows, but they have a cross-over zone between QS=20 and QS=40. The adversarial perturbation dominates the accuracy reduction before the cross-over point (small QS), however, after that, both acc_m (1 pass) and acc_l will decrease due to the

Table 1: The defense efficiency (classification accuracy on adversarial examples) of selected defense methods against different adversarial attacks.

	FGSM	BIM	DeepFool	CW2	CW0	CWi	Average	acc_l	Time (s)
No defense (%)	0	0	11	10	0	0	3.5	69.5	0.11
Bit-depth (5-bit) (%)	2	0	21	68	7	33	21.83	**69.4**	0.04
JPG (90) (%)	5	9	9	68	5	32	21.33	69	0.11
Quilting (%)	48	61	47	50	48	49	50.5	63.5	32.47
TVM (%)	33	42	68	77	49	90	59.8	60	38.89
FD-1P (%)	13	35	63	86	61	78	56	68.5	0.16
FD-2P (%)	**92**	**99**	**99**	**99**	58	**99**	**91**	68.5	0.16

enlarged QS. On the other hand, acc_m (2 pass) increases consistently as QS increases because of additional defensive quantization in compression stage. *Therefore, we set S_2=20 and S_1=30 for the top-n largest $\delta'_{i,j}$ (AS Band) and the others (MD Band), respectively, to better balance the acc_m and acc_l, according to our flow in Fig. 2. Fig. 4 (d) validates that such a configuration of (S_1, S_2) at $n = 15$ minimizes the degradation of acc_l ($\leq 1\%$).*

Theoretical validation of defensive QS. Fig. 4 (e) further compare our analytic results (see Eq. 5) with experimental results for selecting QS. We use FGSM attack with 5 different perturbation strengths (i.e. $\epsilon = 0.005, 0.007, 0.01, 0.012, 0.015$) as an example. The corresponding analytic QS values based on Eq. 5 should be: 20.5, 28.7, 41, 49.2 and 61.44, respectively (dash lines in Fig. 4 (e)). As expected, those analytical values are in excellent agreement with the experimental results when the defense efficiencies reach 100%.

4.3. Enhanced Robustness Against AE

Based on our explorations on parameters optimization in section 4.2, we adopt $S_1 = 30, S_2 = 20, n = 15$ to evaluate the overall defense efficiency. Note although we focus on defense efficiency, the images compressed by our method still provide acceptable visual quality (detailed results are summarized in the supplemental material).

4.3.1 Gray Box Mitigation

Table. 1 compares the defense efficiency of two proposed methods (i.e., 1-pass feature distillation **FD-1P** and 2-pass version **FD-2P**) with five baselines–no defense, JPEG, Bit-depth, Quilting and TVM against 6 selected adversarial examples for MobileNet. *Note that JPEG (90% quality) and Bit-depth (5-bits) are conducted under the premise of $\leq 1\%$ legitimate classification accuracy reduction. However, the other two methods Quilting and TVM, cannot satisfy this constraint, so we compare our approach with those two methods on both acc_l and acc_m.*

Comparison with bit-depth and JPEG. We first limit our comparisons to the defense with $\leq 1\%$ reduction of acc_l under no defense. In this case, Quilting and TVM are not included and will be compared separately.

Overall, FD-2P shows much better performance than that of FD-1P (56% v.s. 91% on average). Compared with no defense baseline, our FD-2P improves the average accuracy on adversarial examples from $\sim 3.5\%$ to $\sim 91\%$, which demonstrates the best mitigation efficiency among all methods. Moreover, FD-2P can significantly outperform two other defensive baselines among all selected adversarial examples, i.e. improved by $\sim 69\%$ (or $\sim 73\%$) on average than the bit-dept (5-bit) or JPEG on both DNN models.

Particularly, for L_∞ attacks like FGSM, BIM and CWi, existing model-agnostic methods show very limited efficiency. Similarly, our one pass method FD-1P shows marginal improvement when compared with the existing approaches. However, our two pass method FD-2P can almost completely remove this type of L_∞ perturbations and deliver the best defense efficiency. Besides, for the L_2 attacks, especially CW_2, existing defense methods show good defense efficiency ($\sim 68\%$). Again FD-2P can rectify this kind of adversarial examples with almost 100%. Compared with L_∞ and L_2, the improvement of L_0 attacks (CW_0) is less significant, however, FD-2P still achieves more than 50% defense efficiency improvement comparing with bit-depth and JPEG. That is because, JPEG (90% quality) uses small quantization steps (or large QFs) to maintain the quality of legitimate images for desirable accuracy, however, is also resulting in a low defense efficiency. Bit-depth roughly quantizes all image pixels uniformly, while our method distills the features in a more fine-grained manner by maximizing the loss of adversarial perturbations and minimizing the distortions of benign features.

Comparison with Quilting/TVM. We also compare our solutions with Quilting and TVM in three aspects: acc_m, acc_l, and processing-time-per-image. Our average defense efficiency is much higher than the other two, i.e. 56%/91% (FD-1P/FD-2P) v.s. 50.5% (quilting), 59.8% (TVM). We also achieve the best testing accuracy (acc_l), that is 68.5% (FD-1P/2P) v.s. 63.5% (quilting), 60% (TVM). Moreover, we improve the processing-time-per-image (i.e., 0.15s on FD-1P) by $\sim 216\times$ ($\sim 259\times$) compared with Quilting (32.4s) and TVM (38.8s), or 0.15s (FD-1P) v.s. 32.4s (quilting) and 38.8s (TVM), as Table 1 shows.

In general, our proposed feature distillation is particu-

Table 2: The defense efficiency (accuracy on adversarial examples–acc_m and accuracy on legitimate images–acc_l on ImageNet, against adaptive adversarial attack–BPDA.

	None	Bit-depth	Quilting	TVM	JPEG(75)	JPEG(20)	JPEG(10)	FD(1x)	FD(2x)	FD(3x)
acc_m (%)	0	0	0	0	0	34	45	10	42	60
acc_l (%)	78	77	72	68	74	68	61	77	76	74

larly effective to mitigate stronger attacks (i.e. CW attacks with least perturbations but $\sim$ 100% attack success rate) crafted from complex datasets like ImageNet. Our solution demonstrates great potentials to safeguard the DNNs against adversarial attacks in practical applications, given that it is likely the attackers prefer to generate stronger adversarial examples with minimum adversarial perturbations from realistic large-scale dataset so as to evade any possible defense methods.

4.3.2 White Box Mitigation

In this section, we evaluate our method against recent BPDA attack, of which adversary knows the defense method and iteratively generates adversarial examples according to the defense. We implement our defense–Feature Distillation (FD-1P) in the released BPDA attack [3] code at GitHub, using the same Inception v3 model and 100 iterations for BPDA. The accuracy of benign examples (adversarial examples) after defense-acc_l (acc_m), for different methods are reported in Table 2.

First, Bit-depth, quilting and TVM does not offer any defense against BPDA, as expected. **Second**, JPEG can slightly mitigate BPDA by degrading image quality, i.e. quality factor from 75 to 10, defense efficiency (acc_m) is improved from 0 to 45%. This is consistent with the recent result [21]. However, acc_l drops by 17% compared to baseline (61% v.s. 78%), which is unacceptable. This reason is because in order to eliminate a large perturbation of BPDA attack in the lowest frequency component in JPEG, a significant large quantization factor (QF) will be needed. As a result, large quantization errors will occur in high frequency components, thereby significantly hurting acc_l. **Third**, On the other hand, our solution can provide the best defense efficiency against BPDA with least acc_l reduction among all solutions, i.e. from FD (1X) to FD (3X), acc_m is improved from 10% to 60%, with merely 1%-3% acc_l reduction compared to original 78%. FD-1$\times$, 2$\times$, 3$\times$ represent the quantization step (QS) of FD adopted in Table. 1 (reference), 2 times and 3 times of the referred QS, respectively.

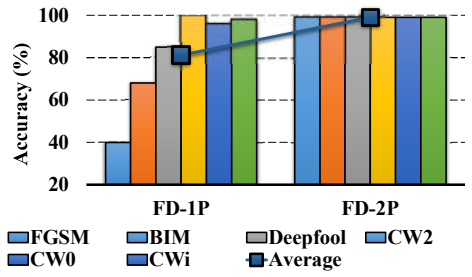

Figure 5: Defense efficiency of black-box setting for different attack and defense mechanisms on ImageNet.

4.3.3 Black Box Mitigation

We follow the work [11] for black-box analysis: DNN model used for testing is trained on transformed dataset (Feature Distillation), while attackers generate adversarial examples from the model trained on the original dataset. The crafted examples have high transferability between the two models for fair black-box analysis. We adopt MobileNet and the results of our methods are shown in Fig. 5.

The average defense efficiency is improved from 56%/91% (Table 1) to 81%/99% (black-box) for our FD-1P/FD-2P method, respectively. These results indicate that our method defends against black-box setting efficiently. This is also consistent with the following conclusion based on [3, 11]: Black box setting shows weak attack efficiency against the input-transformation based defenses.

5. Conclusion

As the robustness of DNN is significantly challenged by a variety of adversarial attacks, existing studies investigate the standard JPEG compression as a defense method, however, it is far from satisfactory in terms of both defense efficiency and testing accuracy. In this work, we propose the DNN-favorable feature distillation method by re-architecting the JPEG compression framework. Compared with existing model-agnostic defense approaches, our "feature distillation" can simultaneously reduce the adversarial attack success rate and maximize the testing accuracy on legitimate examples. Experimental results show that our method can improve the defense efficiency from $\sim$ 20% to $\sim$ 90% over most recent model-agnostic approaches with only marginal accuracy degradation ($\leq$ 1%), while significantly improving the processing time per image ($\sim$ 260$\times$ speedup). Our method also demonstrates the best defense efficiency against latest adaptive attack–BPDA ($\sim$ 60%) with least accuracy drop ($\sim$ 1%) when compared with other input-transformation based defenses.

Acknowledgement

This work is partially supported by the NSF under Grant CNS-1840813.

References

[1] M. Abadi, P. Barham, et al. Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16.

[2] N. Akhtar and A. Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *arXiv preprint arXiv:1801.00553*, 2018.

[3] A. Athalye, N. Carlini, and D. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, July 2018.

[4] A. E. Aydemir, A. Temizel, and T. T. Temizel. The effects of jpeg and jpeg2000 compression on attacks using adversarial examples. *arXiv preprint arXiv:1803.10418*, 2018.

[5] A. N. Bhagoji, D. Cullina, and P. Mittal. Dimensionality reduction as a defense against evasion attacks on machine learning classifiers. *arXiv preprint arXiv:1704.02654*, 2017.

[6] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *Security and Privacy (SP), 2017 IEEE Symposium on*, pages 39–57. IEEE, 2017.

[7] N. Das, M. Shanbhogue, S.-T. Chen, F. Hohman, L. Chen, M. E. Kounavis, and D. H. Chau. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression. *arXiv preprint arXiv:1705.02900*, 2017.

[8] N. Das, M. Shanbhogue, S.-T. Chen, F. Hohman, S. Li, L. Chen, M. E. Kounavis, and D. H. Chau. Shield: Fast, practical defense and vaccination for deep learning using jpeg compression. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 196–204, 2018.

[9] G. K. Dziugaite, Z. Ghahramani, and D. M. Roy. A study of the effect of jpg compression on adversarial images. *arXiv preprint arXiv:1608.00853*, 2016.

[10] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[11] C. Guo, M. Rana, M. Cissé, and L. van der Maaten. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017.

[12] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[13] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *Artificial Intelligence Safety and Security*, pages 99–112, 2018.

[14] F. Liao, M. Liang, Y. Dong, T. Pang, J. Zhu, and X. Hu. Defense against adversarial attacks using high-level representation guided denoiser. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1778–1787, 2018.

[15] Q. Liu, T. Liu, Z. Liu, Y. Wang, Y. Jin, and W. Wen. Security analysis and enhancement of model compressed deep learning systems under adversarial attacks. In *Proceedings of the 23rd Asia and South Pacific Design Automation Conference*, ASPDAC '18, pages 721–726, Piscataway, NJ, USA, 2018. IEEE Press.

[16] S. M. Moosavi Dezfooli, A. Fawzi, and P. Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, number EPFL-CONF-218057, 2016.

[17] E. Ohn-Bar and M. M. Trivedi. Looking at humans in the age of self-driving and highly automated vehicles. *IEEE Transactions on Intelligent Vehicles*, 1(1):90–104, 2016.

[18] A. Prakash, N. Moran, S. Garber, A. DiLillo, and J. Storer. Deflecting adversarial attacks with pixel deflection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8571–8580, 2018.

[19] R. Reininger and J. Gibson. Distributions of the two-dimensional dct coefficients for images. *IEEE Transactions on Communications*, 31(6):835–839, 1983.

[20] P. Samangouei, M. Kabkab, and R. Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. *arXiv preprint arXiv:1805.06605*, 2018.

[21] U. Shaham, J. Garritano, Y. Yamada, E. Weinberger, A. Cloninger, X. Cheng, K. Stanton, and Y. Kluger. Defending against adversarial images using basis functions transformations. *arXiv preprint arXiv:1803.10840*, 2018.

[22] N. Smolyanskiy, A. Kamenev, J. Smith, and S. Birchfield. Toward low-flying autonomous mav trail navigation using deep neural networks for environmental awareness. *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4241–4247, 2017.

[23] Y. Song, T. Kim, S. Nowozin, S. Ermon, and N. Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. *arXiv preprint arXiv:1710.10766*, 2017.

[24] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[25] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. M. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

[26] J. Uesato, B. O'Donoghue, A. v. d. Oord, and P. Kohli. Adversarial risk and the dangers of evaluating against weak attacks. *arXiv preprint arXiv:1802.05666*, 2018.

[27] G. K. Wallace. The jpeg still picture compression standard. *IEEE transactions on consumer electronics*, 38(1):xviii–xxxiv, 1992.

[28] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. Yuille. Mitigating adversarial effects through randomization. *arXiv preprint arXiv:1711.01991*, 2017.

[29] W. Xu, D. Evans, and Y. Qi. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. In *Proceedings of the 2018 Network and Distributed Systems Security Symposium (NDSS)*, 2018.

[30] S. Ye, Q. Sun, and E.-C. Chang. Detecting digital image forgeries by measuring inconsistencies of blocking artifact. In *Multimedia and Expo, 2007 IEEE International Conference on*, pages 12–15. IEEE, 2007.

[31] X. Zhang, S. Wang, K. Gu, W. Lin, S. Ma, and W. Gao. Just-noticeable difference-based perceptual optimization for jpeg compression. *IEEE Signal Processing Letters*, 24(1):96–100, 2017.