# Feature Distillation: DNN-Oriented JPEG Compression Against Adversarial Examples

**Zihao Liu[1], Qi Liu[1], Tao Liu[1], Yanzhi Wang[2], Wujie Wen[1]**
[1] Department of Electrical and Computer Engineering, Flordia International University
[2] Electrical Engineering and Computer Science Department, Syracuse University
zliu021@fiu.edu, qliu020@fiu.edu, tliu023@fiu.edu, ywang393@syr.edu, wwen@fiu.edu



## Abstract

Deep Neural Networks (DNNs) have achieved remarkable performance in a myriad of realistic applications. However, recent studies show that well-trained DNNs can be easily misled by adversarial examples (AE) – the maliciously crafted inputs by introducing small and imperceptible input perturbations. Existing mitigation solutions, such as adversarial training and defensive distillation, suffer from expensive retraining cost and demonstrate marginal robustness improvement against the state-of-the-art attacks like CW family adversarial examples. In this work, we propose a novel low-cost "feature distillation" strategy to purify the adversarial input perturbations of AEs by redesigning the popular image compression framework "JPEG". The proposed "feature distillation" wisely maximizes the malicious feature loss of AE perturbations during image compression while suppressing the distortions of benign features essential for high accurate DNN classification. Experimental results show that our method can drastically reduce the success rate of various state-of-the-art AE attacks by $\sim 60\%$ on average for both CIFAR-10 and ImageNet benchmarks without harming the testing accuracy, outperforming existing solutions like default JPEG compression and "feature squeezing".

## 1 Introduction

Thanks to the recent machine learning model innovation and computing hardware advancement, the past decade has witnessed unprecedented success of deep neural networks (DNNs) across many real world applications such as image recognition, natural language processing, anomaly detection, driver-less cars, drones, etc [Bojarski *et al.*, 2016; Bourzac, 2016; Giusti *et al.*, 2016; Andor *et al.*, 2016; Graves *et al.*, 2013]. However, recent studies have shown that DNN models are inherently vulnerable to adversarial examples (AEs) [Goodfellow *et al.*, 2014; Szegedy *et al.*, 2013], i.e. malicious inputs crafted by adding small and human-imperceptible perturbations to normal and benign inputs, strongly fooling the cognitive function of DNNs such as target misclassification. For example, in image recognition, adversarially manipulating the perceptual systems of autonomous vehicles by physically captured adversarial images, i.e. via camera or sensor [Ohn-Bar and Trivedi, 2016; Smolyanskiy *et al.*, 2017], can lead to the misreading of road signs, thus causing potential disastrous consequences in DNN-based cyber-physical systems.

Many countermeasures have been proposed to enhance the robustness of DNNs against adversarial examples, including DNN model-specific hardening strategies and model-agnostic defenses (or adversarial examples preprocessing techniques) [Guo *et al.*, 2017]. However, these solutions either require expensive computation or show limited success against state-of-the-art attack benchmarks. For example, typical model-specific solutions like "adversarial training" or "defensive distillation" refines the model parameters to defend the AEs or masks the adversarial gradient from generating stronger AEs, but they suffer from high training cost due to iterative retraining procedure. Moreover, "defensive distillation" is proved to be ineffective to counteract most recent Carlini & Wagner attacks (or CW family attacks) [Carlini and Wagner, 2017]. The model-agnostic approaches such as input dimensionality reduction [Bhagoji *et al.*, 2017] or direct JPEG compression [Dziugaite *et al.*, 2016; Das *et al.*, 2017; Guo *et al.*, 2017] appear too simple to sufficiently remove adversarial perturbations from input images without harming the DNN model accuracy [Dziugaite *et al.*, 2016]. "feature squeezing", as one of the most powerful AE detection techniques belonging to model-agnostic approaches, is able to detect AEs with high accuracy and few false positives, however, it is still very complicated because multiple models are needed in order to accurately compare the model's prediction on the original sample with its predictions on the sample after feature squeezing [Xu *et al.*, 2018].

In this work, we focus on improving the effectiveness of JPEG compression based model-agnostic defense against adversarial examples in image classification. As we shall show later, directly deploying standard lossy JPEG compression algorithm as a defense method [Dziugaite *et al.*, 2016; Das *et al.*, 2017] neither effectively removes the adversarial perturbations nor guarantees the accuracy of benign samples. Hence, we for the first redesign the JPEG compression framework to be DNN-favorable (instead of centering around human-visual system (HVS)), and develop a novel low-cost strategy, called *"feature distillation"*, augmented from standard JPEG, to simultaneously improve the DNN robustness against AE attacks while ensuring DNN model's testing accuracy. Our major contributions can be summarized as:

1. We analyze the frequency feature distributions of adversarial perturbations of popular AEs during JPEG compression and propose a semi-analytical method to guide the quantization process to maximize the effectiveness of filtering the adversarial features;
2. We characterize the importance of input features for DNNs by leveraging the statistical frequency component analysis of each image during JPEG compression, and further develop DNN-oriented (rather than HVS) compression method to recover the accuracy reduction of benign samples because of feature loss incurred by purifying AEs at high JPEG compression rates;
3. Experimental results show that our proposed *"feature distillation"* can reduce the success rate of various stat-of-the-art AE attacks $\sim 60\%$ on average with high classification accuracy guarantee, outperforming existing solutions like direct JPEG compression and most recent "feature squeezing".

Our proposed *"feature distillation"* method is built upon the light modifications of widely adopted JPEG compression and does not require any expensive model retraining or multiple model predictions, thus is very low-cost. It can be a black-box defense method since it does not need any knowledge about the model or AEs, and is orthogonal to existing model-specific hardening techniques like "adversarial training" or "gradient masks".

# 2 Background and Motivation

## 2.1 HVS-based JPEG Compression

As the major context to be understood by DNNs, images are usually stored and transferred in the compressed format. Among all compression standards, JPEG [Wallace, 1992] is a popular lossy compression standard for digital images. A typical JPEG compression mainly consists of Image Partition, Discrete Cosine Transformation (DCT), Quantization, Zig-zag reorder and Entropy Coding etc [Wallace, 1992]. The raw image is first partitioned into multiple $8 \times 8$ blocks, followed by a block-wise DCT transform (64-coefficient DCT), which results in 1 Direct Current (DC) coefficient and 63 Alternating Current (AC) coefficients for 64 different frequencies. The DC coefficient denotes the average color of $8 \times 8$ region while the remaining ones (AC) represent color change across the block. *Since Human-Visual System (HVS) is less sensitive to the high frequency components [Zhang* et al.*, 2017], the high (low) frequency coefficients are usually scaled more (less) and then rounded to nearest integers by performing element-wise division based on a pre-characterized* $8 \times 8$ *Quantization Table (Q-Table)* [Wallace, 1992]. Then the quantized results are reordered in a zig-zag order as follows, with the DC coefficient followed by AC coefficients of increasing frequency. In entropy coding, the zig-zag style DC and AC coefficients are coded by differential pulse-code modulation and run-length coding, respectively. Finally, coded results are feeding forward to Huffman or Arithmetic Coding for further compression and eventually assembled as frames of JPEG file. The trade-off between image quality and compression ratio is conducted by scaling each element in Q-Table via a parameter named "Quantization Factor" (QF) [Ye *et al.*, 2007]. A higher compression rate can be achieved through a lower QF (enlarged Q-Table).

The decompression procedure follows the reversed steps of the compression. Note that the Q-Table based quantization is the only irreversible procedure to cause the information loss among all steps.

## 2.2 Adversarial Examples (AEs)

Adversarial examples are maliciously crafted inputs, which are dedicated to misleading the DNN classification by exerting small input perturbations.

**FGSM and BIM.** The fast gradient sign method (FGSM) [Goodfellow *et al.*, 2014] is a fast algorithm to compute perturbations subject to an $L_\infty$ constraint. Each input perturbation can be derived in the direction of the sign of the gradient of the loss function:

$$X^{'} = X + \epsilon \cdot sign(\nabla_X L(X, y_{true})) \tag{1}$$

where $X^{'}$ is the polluted input, $L$ is the loss for a specific DNN model at training phase, $y_{true}$ is the correct label for input $X$ and $\epsilon$ is the perturbation strength. To further enhance attacking strength, the basic iterative method (BIM) augmented from FGSM is also proposed [Kurakin *et al.*, 2016]. It adds a small perturbation $\alpha$ in each iteration until reaches $\epsilon$ or achieves successful attack:

$$X_0^{'} = X,\ X_i^{'} = X_{i-1}^{'} + clip_\epsilon\left(\alpha \cdot sign(\nabla_X L(X, y_{true}))\right) \tag{2}$$

where the clipping equation, $clip_\epsilon(n)$, performs clipping on each pixel when it reaches $\epsilon$. It is worth noting that FGSM (or BIM) is a computationally efficient fast attack rather than an optimal attack with minimal adversarial perturbations [Kurakin *et al.*, 2016].

**Deepfool.** Deepfool [Moosavi Dezfooli *et al.*, 2016] uses geometrical knowledge to compute the $L_2$ distance between input $X$ and decision boundary $P$. In this method, the DNN is treated by the adversarial agent as a linear classifier and each class is separated by a hyperplane. The algorithm finds the nearest hyperplane from $X$, and uses geometrical knowledge to calculate the projection distance. Since DNN function is not strictly linear, deriving an effective adversarial example usually requires multiple iterations. The optimized perturbations can be smaller than that of FGSM, thus are more difficult for humans to detect.

**The Carlini & Wagner (CW) Method.** Carlini and Wagner [Carlini and Wagner, 2017] have proposed three methods to craft adversarial examples based on $L_0, L_2$ and $L_\infty$ norms as distance metrics. The adversarial agent searches for $w$ that is a new variable introduced by authors.

$$minimize\left[\|\tfrac{1}{2}\left(tanh(w)+1\right) - X\|_p^2 + c \cdot f\left(\tfrac{1}{2}(tanh(w)+1)\right)\right] \tag{3}$$

where $f(\cdot)$ is the objective function designed by authors based on the loss function:

$$f(x) = max\left(max\left\{Z(X)_i \;\; i \neq t\right\} - Z(X)_t, -\kappa\right) \tag{4}$$

Here $Z(X)_i$ is the output of pre-softmax "logit" for class $i$. We can adjust $\kappa$ to control the confidence of adversarial attack

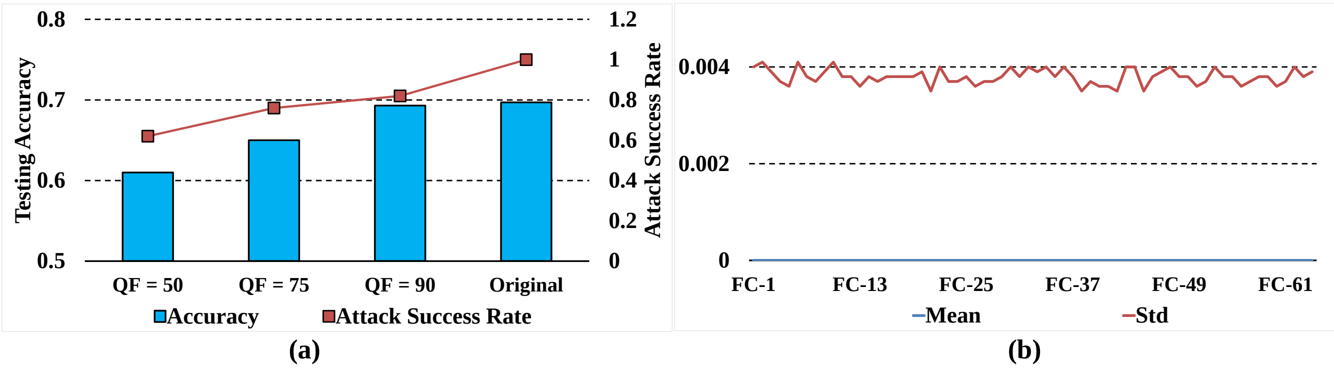


Figure 1: (a) Testing accuracy v.s. attack success rate at different QFs of JPEG; (b) Statistical information of FGSM-based AE perturbations in frequency domain (FC denotes frequency component)

success. Note that CW family attacks are the most recent and strongest attacks with least total distortion but can succeed in finding AEs for 100% of images on defensively distilled networks [Papernot *et al.*, 2016]. It is the state-of-the-art benchmark to evaluate the effectiveness of any defense attempts.

### 2.3 Accuracy and AE Defense Efficiency of JPEG

*DNN suffers from both testing accuracy loss and weak AE defense efficiency if directly employing existing HVS-based compression techniques*: To explore how existing compressions can impact the testing accuracy of DNNs, we have conducted the following experiment: training DNN model by high quality JPEG images (QF=100), and testing it with images at various QFs (i.e., QF=100, 90, 75, 50); A representative DNN example–"MobileNet" [Howard *et al.*, 2017] is trained with the ImageNet dataset for large scale visual recognition. Also we take FGSM as an example to explore the AE defense efficiency at various selected QFs. As Fig. 1 (a) shows, the "top-1" testing accuracies degrade significantly as the compression ratio increases (i.e., QF from 100 to 50), however, the AE defense efficiency increases. To achieve the best defense efficiency (attack success rate = 0.62 at QF = 50), the accuracy is even degraded by $\sim 8\%$ on benign images than that of the original one (QF=100). On the other hand, JPEG can only guarantee the testing accuracy of DNN models at quite high QF values, but such less compressed JPEG cannot defend AE attacks effectively.

The next question is why default JPEG shows weak defense efficiency for AE attacks. Again We take the FGSM based AE to explore the perturbation distribution in the frequency domain. By following the JPEG compression procedure, we transfer the malicious perturbations to $8 \times 8 = 64$ frequency components and analyze the corresponding statistical information. As Fig. 1 (b) shows, the means and standard deviations of the DCT coefficients of malicious perturbation at all 64 frequency components are almost the same. Therefore, the HVS-based JPEG compression, i.e., compressing more (less) on high (low) frequency components of an input image, is neither suitable for defending AE attacks nor for achieving high compression rate without accuracy loss.

## 3 Our Approach

Leveraging the JPEG compression technology as a defense method against AE attacks has been studied in previous works [Dziugaite *et al.*, 2016; Das *et al.*, 2017; Guo *et al.*, 2017], however, none of them have explained in details the defense principles and how to further optimize. In this section, we first provide a detailed analysis on how to utilize compression to minimize AE attack success rate. As this lossy compression will still reduce the classification accuracy as described in Section 2.3, we then develop a **DNN-oriented JPEG compression** method to compensate the reduced accuracy. The target of the proposed framework is to overcome both AE attacks and the accuracy reduction issue.

### 3.1 Analysis of Compression for Mitigating AEs

We propose to use spectral filter in JPEG compression on DNN inputs (images), in order to mitigate the adversarial perturbations. The inputs with adversarial perturbations will be placed into "DCT transformation" and "quantization" processes in JPEG compression.

Assuming for each $8 \times 8$ block in the input image $I$, perturbation $E$ is added to craft the adversarial example $I+E$, which can be generated through AE algorithms such as FGSM with perturbation intensity represented as $\epsilon$. In JPEG compression, the step of DCT transformation will project the image from spatial domain to spectral domain by Discrete Cosine Transform (DCT), which is essentially a linear operation. As a result, the original input and adversarial perturbations could be linearly separated as:

$$DCT(I+E) = DCT(I) + DCT(E) = C_I \cdot B + C_E \cdot B \quad (5)$$

where $C_I$ and $C_E$ are the DCT coefficients of $I$ and $E$, respectively, for the $8 \times 8$ image block, and $B$ is the DCT transformation basis. Typically, following the DCT transformation, the maximum magnitude of $C_E$ can be calculated by the summation of all the 64 frequency components and each term is bounded by $cos(\theta) \cdot \epsilon$. Thus we have $-8 \cdot \epsilon < C_E < 8 \cdot \epsilon$. Furthermore, the DCT coefficients will be quantized in JPEG compression process, providing a good opportunity for filtering the adversarial perturbations in the spectral domain. The quantization in JPEG compression can be approximated as:

$$Round\left({}^{C_I}/_{Q} + {}^{C_E}/_{Q}\right) \approx Round\left({}^{C_I}/_{Q}\right) + Round\left({}^{C_E}/_{Q}\right) \quad (6)$$

where $Q$ is the quantization step (QS) size. To completely eliminate the perturbation, an appropriate $Q$ that satisfies the following equation should be selected:

$$Round\left({}^{C_E}/_{Q}\right) = 0 \Rightarrow Q > 2\left|C_E\right|, \; C_E \in (-8\epsilon < C_E < 8\epsilon) \quad (7)$$

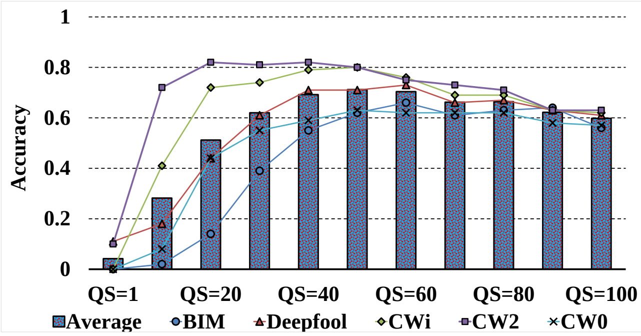


Figure 2: Quantization step vs. accuracy for various AEs.

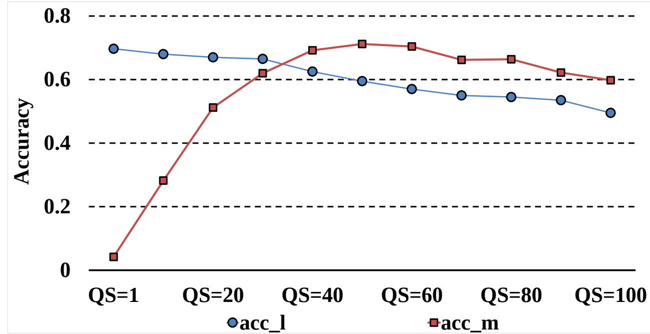


Figure 3: The impact of quantization steps on $acc_l$ and $acc_m$.

To this end, as long as the QS size $Q > 16 \cdot \epsilon$, the $C_E$ perturbations will be properly filtered.

We evaluate the effectiveness of our method in defending against AE attacks, under different QS settings with various types of AEs. The ImageNet dataset and a pre-trained DNN model–MobileNet [Howard *et al.*, 2017], are adopted. All the frequency components are set to have the same quantization step. As Fig. 2 shows, for all types of AE attacks, including recent CW family attacks, the effectiveness of defending against malicious inputs increases as QS grows. Furthermore, the defending effectiveness of all AE attacks reach a similar level when QS is sufficiently large. In this case, all the $C_E$ coefficients have been zeroing out after a customized quantization (i.e. $Q \sim 40$), rather than the default JPEG quantization value, completely eliminating the adversarial perturbations. However, further enlarging QS can lead to lower accuracy due to the over-quantized non-malicious features.

Therefore, there still exists one important challenge for this conceptual defense technique, since the quantization approach can also affect the overall DNN accuracy due to the loss of benign image features, i.e. $C_I$ (see Eq.6), in the JPEG compression procedure. In practice, the overall accuracy of a trained DNN model can be expressed as:

$$P = P_l \cdot acc_l + P_m \cdot acc_m \tag{8}$$

where $P_l$ and $P_m$ are the probabilities of legal and malicious inputs, respectively; $acc_l$ and $acc_m$ are the predication accuracies of legal and malicious inputs, respectively.

Although the $P_l$ and $P_m$ are unknown in realistic scenario, we intend to boost $P$ by maximizing both $acc_l$ and $acc_m$. A high $acc_l$ requires a small quantization step to keep the benign features as much as possible; on the contrary, a high $acc_m$ needs a large quantization step to eliminate AE perturbation. Fig. 3 illustrates the impact of quantization step $QS$ on $acc_l$ and $acc_m$. Here $acc_m$ is the average accuracy over all types of AE attacks with a specific $QS$ value. As Fig. 3 shows, the two curves at first demonstrate an opposite trend and have a cross-over at $QS = 30 - 40$. Before the cross-over, malicious perturbation dominates the accuracy reduction, but after the cross-over, both accuracies decrease as the $QS$ increases. A desirable value such as $QS = 40$ can provide good enough defense effectiveness against AE attacks. However, the 7.5% decrease of $acc_l$ is unacceptable for the benign inputs.

## 3.2 DNN-Oriented JPEG Compression Method

In order to boost the testing accuracy discussed in section 3.1, we develop a DNN-oriented JPEG compression method by redesigning the quantization table. In this section, we first provide an analysis of the difference between HVS and DNN on feature extractions, then propose an effective re-design of the quantization table for mitigating accuracy loss.

**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 $8 \times 8$ DCT in JPEG compression:

$$x_k = \sum_{i=0}^{i=7} \sum_{j=0}^{n=7} c_{(k,i,j)} \cdot b_{(i,j)} \tag{9}$$

where $c_{(k,i,j)}$ and $b_{(i,j)}$ are the DCT coefficient and corresponding 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:

$$\partial F / \partial b_{(i,j)} = \partial F / \partial x_k \times \partial x_k / \partial b_{i,j} = \partial F / \partial x_k \times c_{(k,i,j)} \tag{10}$$

Eq. (10) implies that the contribution of a frequency component ($b_{i,j}$) to the DNN result will be mainly decided by its associated DCT coefficient ($c_{(k,i,j)}$) and the importance of the pixel ($\partial F / \partial x_k$). Here $\partial F / \partial x_k$ is obtained after the DNN training, while $c_{(k,i,j)}$ 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 only focuses on compressing more on the less sensitive frequency components to HVS. As a result, in order to defend AE attacks, aggressive compression is required, making DNNs easily misclassified if the original versions contain important high frequency features. The DNN models trained

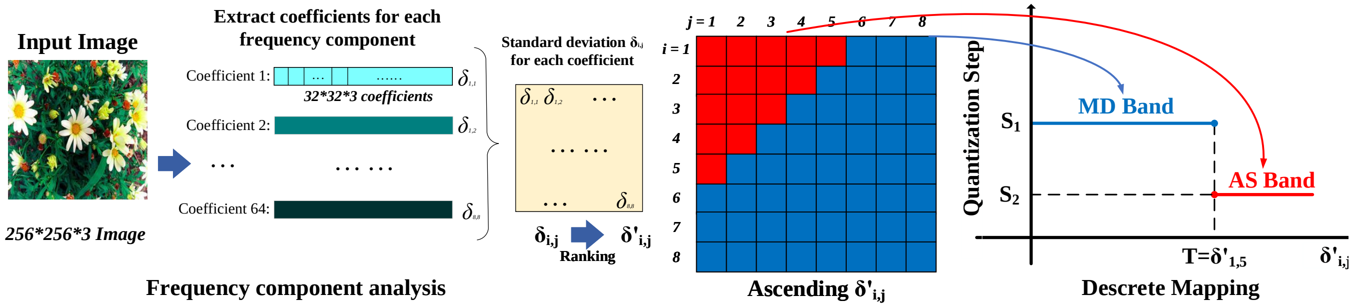


Figure 4: An overview of heuristic design flow of DNN-Oriented JPEG Compression Method.

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)).

The proposed "feature distillation" AE defense method is developed upon a heuristic design flow (see Fig. 4): 1) characterize the importance of each frequency component through frequency analysis on the testing images; 2) reduce the quantization step of the most sensitive frequency components based on its statistical information and guarantee the testing accuracy.

**Frequency Component Analysis.** For each input testing 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 [Reininger and Gibson, 1983]. Prior works [Reininger and Gibson, 1983] have proven 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. The detailed procedures can be summarized as: Each image will be first partitioned into $N$ $8 \times 8$ 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.

**Quantization Table Design.** Once the importance of frequency components is identified based on the standard deviations of DCT coefficients ($\delta_{i,j}$), our next step is to boost the accuracy $\{acc_l\}$. The basic idea is to utilize finer quantizations at the important components by leveraging the intrinsic error resilience property of DNNs. Our analysis in Section 3.1 indicates that a proper selection of $QS = 40$ can effectively mitigate AE perturbations, whereas larger $QS$ will induce more quantization error.

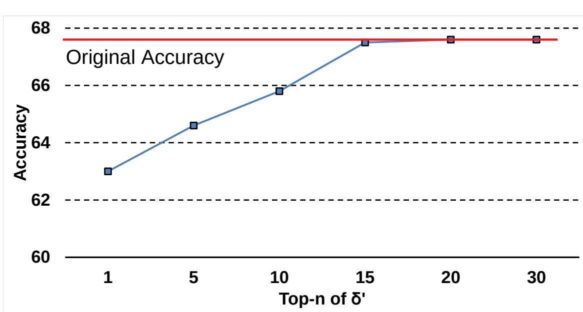


Figure 5: Impact of quantization step on $acc_l$ and $acc_m$ .

Therefore, we remove the quantization errors of the most sensitive frequency components to enhance the testing accuracy by lowering their corresponding quantization steps within the quantization table. Note the table composed of the same QS values ($QS = 40$) delivers the best AE defense efficiency. To maximize the testing accuracy without impacting the AE defense efficiency, the quantization errors for only most important frequency components (but as few as possible) will be eliminated. Specifically, we first sort the magnitude of $\delta_{i,j}$ in an ascending order as $\delta'_{i,j}$, then set the $QS = 1$ and $QS = 40$ for $top - n$ largest $\delta'_{i,j}$ and the remained $\delta'_{i,j}s$, respectively. The relationship between $n$ and $acc_l$ shown in Fig. 5 indicates that set the $QS = 1$ at the position of the top 15 frequency components can ensure the testing accuracy.

To simplify our design, we introduce a discrete mapping function (DM) to derive the quantization step of each frequency band from the associated standard deviation (see Fig. 4):

$$Q_{i,j} = \begin{cases} S_1 & \delta_{i,j} \leq T \\ S_2 & \delta_{i,j} > T \end{cases} \tag{11}$$

where $Q_{i,j}$ is the quantization step at the frequency band $(i, j)$. $T$ is the threshold to categorize the 64 frequency bands according to ascending order of the magnitude of $\delta_{i,j}$. As right part of Fig. 4 shows, the 64 frequency components are divided into two bands: the red colored *accuracy sensitive (AS) band*, consists of 15 largest $\delta'_{i,j}$; the blue colored *malicious defense (MD) band*, consists of the others. Hence, we adopt $T = \delta'_{1,5}$, $S_1 = 40$, and $S_2 = 1$ in our design.

# 4 Evaluation

In this section, we evaluate the robustness against adversarial perturbations of "feature distillation" with the constraint of high classification accuracy on legitimate inputs, since any practical defense approach should well handle malicious samples but at the same time not impact the accuracy of legitimate ones given that both types of data will arrive at a realistic DNN testing.

**Experimental Setup.** Our experiments are conducted on the Tensorflow machine learning framework [Abadi *et al.*,

Table 1: Evaluation results of attacks

| AE | CIFAR-10 | | | | ImageNet | | | |
|---|---|---|---|---|---|---|---|---|
| | Success Rate | $\mathbf{L}_\infty$ | $\mathbf{L}_2$ | $\mathbf{L}_0$ | Success Rate | $L_\infty$ | $L_2$ | $L_0$ |
| **BIM** | 92 | 0.008 | 0.368 | 0.993 | 100 | 0.004 | 1.406 | 0.984 |
| **Deepfool** | 98 | 0.028 | 0.235 | 0.995 | 89 | 0.027 | 0.726 | 0.984 |
| **CW$_2$** | 100 | 0.034 | 0.288 | 0.768 | 90 | 0.019 | 0.666 | 0.323 |
| **CW$_0$** | 100 | 0.650 | 2.103 | 0.019 | 100 | 0.898 | 6.825 | 0.003 |
| **CW$_\infty$** | 100 | 0.012 | 0.446 | 0.990 | 99 | 0.006 | 1.312 | 0.850 |

2016], running with one Intel(R) Xeon(R) 3.5GHz CPU and two GeForce GTX 1080Ti GPUs. Our proposed "feature distillation" method is implemented on the heavily modified EvadeML-Zoo [Xu *et al.*, 2018], a benchmarking and visualization tool for adversarial machine learning.

Two popular image classification datasets are selected in our experiment: small-scale "CIFAR-10" and large-scale "ImageNet". To be consistent with the DNN models used by [Xu *et al.*, 2018], "DenseNet" and "MobileNet" are adopted for training (testing) the datasets–"CIFAR-10" and "ImageNet", respectively. We assume the attacker has the full knowledge on target DNN models. Table 1 summarizes the success rates and distortion measurements of all selected AE candidates, including most recent CW family attacks–$L_0$, $L_2$ and $L_\infty$. The seed images for adding AE perturbations are selected from the first 100 correctly predicted examples in the test (or validation) set from each dataset for all the attack methods. The defense efficiency is measured by the classification accuracy of 100 polluted images after applying the defense method. The legitimate examples classification accuracy is the testing accuracy of benign images processed by the defense method. Two other defense methods, i.e. default JPEG [Dziugaite *et al.*, 2016; Kurakin *et al.*, 2016] and "feature squeezing" [Xu *et al.*, 2018], are selected as the baselines. For a fair comparison, we first *restrict all the defense methods to have the same legitimate classification accuracy* (i.e. less than $1\%$ accuracy reduction than testing accuracy of the images without AEs–$94.84\%$ for CIFAR-10 and $68.36\%$ for ImageNet). Hence, we choose $QF = 90$ in default JPEG method and 5-bit precision on each pixel for "feature squeezing" [Xu *et al.*, 2018].

**Robustness Evaluation.** Fig. 6 and Fig. 7 compare the defense efficiencies of our method with three baselines across various AE attacks on CIFAR-10 and ImageNet, respectively. Compared with the "original" designs without applying any defense mechanisms, our "feature distillation" method improves the average accuracy on adversarial examples from $2\%$ to $63\%$ ($4\%$ to $57\%$) on CIFAR-10 (ImageNet). Our "feature distillation" method achieves the best defense efficiency among all selected AE candidates, i.e. improved by $31\%$ ($12\%$) and $29\%$ ($32\%$) on average than feature squeezing (default JPEG) on CIFAR-10 and ImageNet, respectively. Specifically, our method achieves more than $50\%$ defense efficiency improvement than all two other defense methods for the CW attack–$CW_0$ on the large-scale ImageNet dataset. The major reason is because, the default JPEG needs to use small quantization steps (or large QFs) to maintain the quality of legitimate images for desirable accuracy, however, it also results in a low defense efficiency. "feature squeezing" 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.

Moreover, our proposed "feature distillation" is particularly effective to mitigate the strongest attacks (i.e. CW attacks with least perturbations but $\sim 100\%$ AE 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 strongest AEs with minimum adversarial perturbations from realistic large-scale dataset in order to evade any possible defense methods.

## 5 Conclusion

The robustness of modern DNNs is significantly challenged by various types of AE attacks. Previous works directly employ JPEG compression as a defense method, however, it is not an optimized solution in terms of defense effectiveness and model accuracy. Therefore, we propose a low-cost "feature distillation" method by re-architecting the JPEG compression framework, to achieve remarkable AE defense effectiveness improvement without harming the legitimate image testing accuracy when compared with other defense solutions. The proposed "feature distillation" can simultaneously reduce the AE attack success rate and maximize the legitimate testing accuracy. Experimental results show that our proposed method can reduce the attack success rate $\sim 60\%$ on average at both CIFAR-10 and ImageNet datasets.

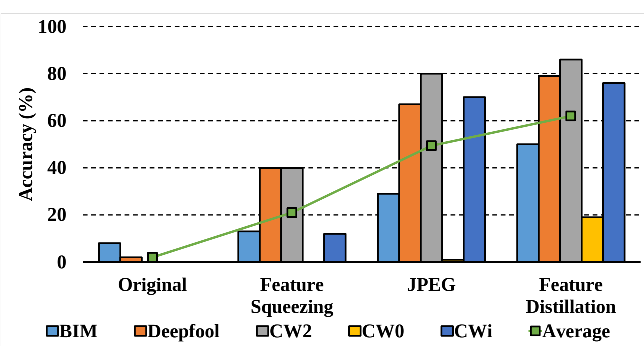


Figure 6: Defense efficiency (accuracy on adversarial examples) on the CIFAR-10 dataset, for different attack and defense mechanisms.

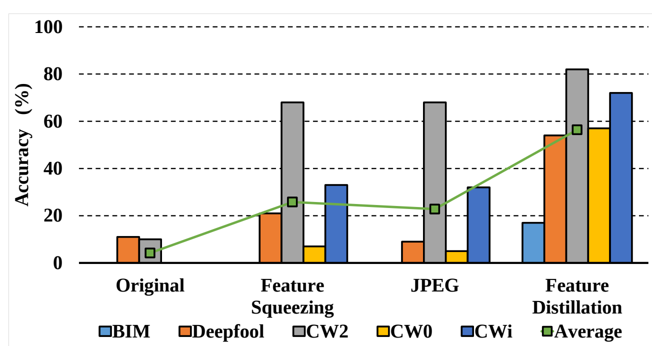


Figure 7: Defense efficiency (accuracy on adversarial examples) on the Imagenet dataset, for different attack and defense mechanisms.

## References


[Abadi *et al.*, 2016] Martín Abadi, Paul Barham, et al. Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

[Andor *et al.*, 2016] Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. Globally normalized transition-based neural networks. *arXiv preprint arXiv:1603.06042*, 2016.

[Bhagoji *et al.*, 2017] Arjun Nitin Bhagoji, Daniel Cullina, and Prateek Mittal. Dimensionality reduction as a defense against evasion attacks on machine learning classifiers. *arXiv preprint arXiv:1704.02654*, 2017.

[Bojarski *et al.*, 2016] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

[Bourzac, 2016] Katherine Bourzac. Bringing big neural networks to self-driving cars, smartphones, and drones. *IEEE Spectrum*, 2016.

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

[Das *et al.*, 2017] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Li Chen, Michael E Kounavis, and Duen Horng Chau. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression. *arXiv preprint arXiv:1705.02900*, 2017.

[Dziugaite *et al.*, 2016] Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M Roy. A study of the effect of jpg compression on adversarial images. *arXiv preprint arXiv:1608.00853*, 2016.

[Giusti *et al.*, 2016] Alessandro Giusti, Jérôme Guzzi, et al. A machine learning approach to visual perception of forest trails for mobile robots. *IEEE Robotics and Automation Letters*, 1(2):661–667, 2016.

[Goodfellow *et al.*, 2014] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[Graves *et al.*, 2013] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE, 2013.

[Guo *et al.*, 2017] Chuan Guo, Mayank Rana, Moustapha Cissé, and Laurens van der Maaten. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017.

[Howard *et al.*, 2017] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[Kurakin *et al.*, 2016] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

[Moosavi Dezfooli *et al.*, 2016] Seyed Mohsen Moosavi Dezfooli, Alhussein Fawzi, and Pascal 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.

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

[Papernot *et al.*, 2016] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *Security and Privacy (SP), 2016 IEEE Symposium on*, pages 582–597. IEEE, 2016.

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

[Smolyanskiy *et al.*, 2017] Nikolai Smolyanskiy, Alexey Kamenev, Jeffrey Smith, and Stan Birchfield. Toward low-flying autonomous mav trail navigation using deep neural networks for environmental awareness. *arXiv preprint arXiv:1705.02550*, 2017.

[Szegedy *et al.*, 2013] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

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

[Xu *et al.*, 2018] Weilin Xu, David Evans, and Yanjun Qi. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. In *Proceedings of the 2018 Network and Distributed Systems Security Symposium (NDSS)*, 2018.

[Ye *et al.*, 2007] Shuiming Ye, Qibin Sun, and Ee-Chien 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.

[Zhang *et al.*, 2017] Xinfeng Zhang, Shiqi Wang, Ke Gu, Weisi Lin, Siwei Ma, and Wen Gao. Just-noticeable difference-based perceptual optimization for jpeg compression. *IEEE Signal Processing Letters*, 24(1):96–100, 2017.