Build a Computationally Efficient Strong Defense Against Adversarial Example Attacks

Changwei Liu, Louis DiValentin, Aolin Ding and Malek Ben Salem

Accenture Cyber Labs, 1201 Wilson Blvd, Arlington, VA, U.S.A.

Keywords: Adversarial Example Attack, Input Transformation Ensembles, Adversarial Example Defense.

Abstract: Input transformation techniques have been proposed to defend against adversarial example attacks in image-classification systems. However, recent works have shown that, although input transformations and augmentations to adversarial samples can prevent unsophisticated adversarial example attacks, adaptive attackers can modify their optimization functions to subvert these defenses. Previous research, especially BaRT (Raff et al., 2019), has suggested building a strong defense by stochastically combining a large number of even individually weak defenses into a single barrage of randomized transformations, which subsequently increases the cost of searching the input space to levels that are not easily computationally feasible for adaptive attacks. While this research took approaches to randomly select input transformations that have different transformation effects to form a strong defense, a thorough evaluation of using well-known state-of-the-art attacks with extensive combinations has not been performed. Therefore, it is still unclear whether employing a large barrage of randomly combined input transformations ensures a robust defense. To answer these questions, we evaluated BaRT work by using a large number (33) of input transformation techniques. Contrary to BaRT’s recommendation of using five randomly combined input transformations, our findings indicate that this approach does not consistently provide robust defense against strong attacks like the PGD attack. As an improvement, we identify different combinations that only use three strong input transformations but can still provide a resilient defense.

1 INTRODUCTION

Machine learning (ML) models, including deep neural networks (DNN), have been successfully applied to a wide range of computer vision tasks (Zhang et al., 2018; Ding et al., 2021; Tang et al., 2020; Ding et al., 2023a; Zang et al., 2022; Ma et al., 2023; Liu et al., 2020). Given the ubiquity of machine learning applications, the security aspects of machine learning models have become increasingly important (Tang et al., 2024). However, studies have shown that attackers can use adversarial examples, the samples of input data slightly modified using an optimization procedure, to cause the misclassification of machine learning models (Szegedy et al., 2013). This raises serious concerns about the security of machine learning models in many real-world applications (Ding, 2022).

Developing strong defenses against adversarial examples has been an important topic. While many other techniques exist, a current focus is on model-agnostic techniques, aiming to remove the adversarial input perturbations from the input through different techniques of transforming the input (we call it input transformation in this paper) (Guo et al., 2017). Researchers have not only explored the robustness of different single input transformation techniques, but also proposed to use the ensemble of input transformations that can provide a stronger defense. Raff et al. (Raff et al., 2019) showed that a computationally stronger defense can be built by stochastically combining a large number of individually input transformation defenses to form a series of input transformation ensembles, even defeating some adaptive at-
tacks by increasing the computational cost of successful adversarial examples to infeasible levels. However, this method has several drawbacks. First, it trades off an increased inference run-time for classification accuracy as each additional transformation is added to the ensemble. Second, it provides no guarantee that the current transformation combination is effective against a strong attack such as EoT attacks (Sitawarin et al., 2022). Third, as shown in Figure 1, the semantic value of an image dramatically changes when multiple input transformations are used upon the image. We have also seen the research efforts that aim to improve the model robustness using adversarial training (Tang et al., 2022). However, adversarial training is not computationally efficient on large and complex datasets, and the model robustness is not effective for larger image perturbations (Shafahi et al., 2020). Using input transformations ensembles against adversarial examples remains an effective method to enterprise users, because it can easily be introduced into a Machine Learning as a Service (MLaaS) pipeline without large architectural changes (Ding et al., 2023b).

To assess the effectiveness of diverse input transformation ensembles in enhancing defense capabilities while minimizing computational expenses and preserving image semantics, we collected 33 input transformation techniques published recently. Subsequently, we conducted a comprehensive evaluation of these ensembles, following the methodology outlined in BaRT. pinpointing those ensembles that deliver robust defense at a reduced number of transformations. The attacks we used to assess input transformations and their ensembles include both state-of-the-art white-box attacks and adaptive attacks designed to evade input transformation techniques.

Overall, our contributions are as follows:
• We rigorously assessed the effectiveness and robustness of 33 input transformation techniques proposed in recent studies. This evaluation involves testing their performance against various adversarial examples generated through white-box and adaptive attacks on CIFAR-10 and ImageNet image datasets.
• We designed and implemented an automated framework to empirically evaluate BaRT’s approach, which advocates for building a robust defense strategy by using a barrage of randomly combined input transformation techniques.
• We analyzed the effectiveness of each combination of transformations and advanced the work by providing insights and recommendations for constructing a computationally efficient but strong defense against adversarial examples. Our contribution is using three strong input transformation ensembles.

The rest of the paper is organized as follows. Section 2 provides background knowledge and related work. Section 3 outlines the implementation of code and the experimental setup used to evaluate input transformations against adversarial examples. In Section 4, we present our analysis of experimental results and engage in a discussion on how we established a computationally efficient but strong defense by identifying ensembles comprising three robust input transformations. Section 5 concludes the paper.

2 BACKGROUND KNOWLEDGE AND RELATED WORK

This section provides the background knowledge and related work of adversarial examples and input transformations.

2.1 Adversarial Examples

Adversarial examples are inputs algorithmically generated by attackers’ applying small but intentionally worst-case perturbations to examples from an image dataset, so that a machine learning model can misclassify the perturbed images. Existing adversarial attacks can be categorized into white-box and blackbox attacks. While, in a white-box attack, an adversary has full knowledge of the target model, including the model architecture and parameters, in a black-box attack, the adversary can only resort to query accesses to generate adversarial samples. In addition, a white-box attack is considered an adaptive attack if the attacker is aware of the defense methods and adapts the attack accordingly (He et al., 2017).

Adversarial examples can be targeted and untargeted. While The targeted attacks are the attacks misguiding the model to a particular class other than the true class, the untargeted attacks are the attacks misguiding the model to predict any of the incorrect classes. Besides, there are four distance metrics, $L_0$, $L_1$, $L_2$, or $L_\infty$, denoting how close an adversarial example needs to be to the original image so that it can keep its semantic value to “fool” a human observer.

2.2 Existing Methods for Generating Adversarial Examples

(Szegedy et al., 2013) discovered that machine learning models are vulnerable to adversarial examples,
other researchers have extensively studied the approaches to generating adversarial examples. Goodfellow et al. proposed a Fast Gradient Sign Method (FGSM), a typical one-step attack algorithm to inject noise into a benign image to cause input misclassification. (Kurakin et al., 2018) extended FGSM to a multi-step attack algorithm named as Basic Iterative Method (BIM) by applying FGSM multiple times with a small step size and clipping pixel values of the intermediate results after each step. As a variant of BIM, (Madały et al., 2017) proposed to constrain the adversarial perturbations by projecting the adversarial sample learned from each iteration into the $L_\infty$-neighborhood of a benign sample. All three attacks, FGSM, BIM and PGD, are untargeted attacks. To achieve a specific targeted adversarial goal, (Papernot et al., 2016) proposed using Jacobian Saliency Map Approach (JSMA) to compute a direct mapping from the input to the output. This approach uses $L_0$ form and is a targeted attack. Using $L_0$, $L_2$ and $L_\infty$ norms, Carlini and Wagner introduced three gradient descent based targeted attacks that have more effective adversarial success rates than previously known adversarial attacks (Carlini and Wagner, 2017b). All the above attacks are white-box attacks, which rely on detailed model information including the gradient of the loss with regard to the input. To have an attack that is applicable to real-world black-box models, Brendel et al. proposed a computationally expensive decision-based adversarial attack (Brendel et al., 2017), in which the algorithm starts from an adversarial example $x_{\text{adv}}$, and then performs random walks toward the boundary between the adversarial and non-adversarial images such that the distance $L$ is minimized.

### 2.3 Input Transformation Techniques and Related Work

Defense against adversarial attacks is broadly classified into proactive (e.g., adversarial training, additional regularization) and reactive (e.g., input transformation, gradient masking) methods. While proactive defenses enhance DNN model robustness, reactive defenses identify adversarial examples in model inputs (Wang et al., 2020). As a reactive defense method, input transformations exploit the observation that small transformations to adversarial attack inputs can often recover the desired classification. Because they are relatively easy to be introduced into machine learning pipelines without large architectural changes, input transformations are appealing as a solution to adversarial examples.

Researchers proposed different input transformation techniques against adversarial examples. Feature squeezing, including color bit depth reduction and spatial smoothing, was suggested by Xu et al. (Xu et al., 2017) to detect adversarial examples. Xie et al. (Xie et al., 2017) used random padding that pads zeros around the input images to defend against adversarial examples. (Prakash et al., 2018) combined two novel techniques, including pixel deflection that randomly replaces some pixels with selected pixels from a small neighborhood and adaptive soft-thresholding that smooths adversarially-perturbed images, to reduce the effects of attacks. (Luo and Pfister, 2018) constructed a Variational Autoencoder (VAE) that maps a high-dimensional feature vector to a lower-dimensional latent vector and then incorporates randomness before mapping it back to the original feature space to defeat adversarial examples.

Many of these defenses have subsequently been broken by adaptive attacks in the white-box threat models. These methods include incorporating the input transformation defense into the adversary’s search procedure (Carlini and Wagner, 2017a) or approximating the obfuscated gradients that have been masked to make it hard for the adversary to generate an adversarial example (Athalye et al., 2018). Even for those defenses used in conjunction with each other, (He et al., 2017) argued that two combined defenses still have a large search space to find an adversarial example that fits the adaptive constraints. To enhance the robustness of defenses, Raff et al. showed it is possible to construct a “computationally” strong defense if the number of single input transformation defenses is large and the combination is created in a randomized fashion (RT defense) (Raff et al., 2019). However, this method comes at a cost of an increased runtime and dramatically changes the image’s semantic value as the number of the combined input transformations is suggested larger than 4. Also, Sitawarin et al. (Sitawarin et al., 2022) argued that the adaptive attack, Backward Pass Differentiable Approximation (BPDA), used to test the RT defense in Raff’s work (Raff et al., 2019) is not sufficiently strong. Thus, the RT defense composed of 5 input transformations, as suggested in (Raff et al., 2019) does not necessarily provide a good defense against a strong start-of-the-art attack.

### 3 EXPERIMENTAL SETUP

In this section, we describe our experimental setup for evaluating the BaRT’s approach of using random input transformation ensembles to construct a strong defense.
Table 1: The 33 Input Transformations in 9 Groups.

<table>
<thead>
<tr>
<th>Group</th>
<th>Input Transformation</th>
<th>Abbreviation in Our Experiments (Section 4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Color precision reduction</td>
<td>Color reduction</td>
<td>color_reduction</td>
</tr>
<tr>
<td>Noise injection</td>
<td>JPEG, Gaussian, Poisson, Speckle, Salt, Pepper and Salt, Pepper</td>
<td>jpeg, noise_gaussian, noise_poisson, noise_speckle, noise_salt, noise_sp, noise_pepper</td>
</tr>
<tr>
<td>Swirl</td>
<td>Swirl</td>
<td>swirl</td>
</tr>
<tr>
<td>Fast Fourier Transform (FFT) perturbation</td>
<td>FFT Perturbation</td>
<td>fft</td>
</tr>
<tr>
<td>Zoom</td>
<td>Random zoom, Random padding, Seam Carving expansion</td>
<td>rescale, padding, seam</td>
</tr>
<tr>
<td>Color space</td>
<td>HSV, XYZ, LAB, YUV</td>
<td>hsv, xyz, lab, yuv</td>
</tr>
<tr>
<td>Contrast</td>
<td>Histogram equalization, Adaptive histogram equalization, Contrast stretching</td>
<td>equalize, adapt_equalize, contrast_stretch</td>
</tr>
<tr>
<td>Grey scale</td>
<td>Grey scale mix, Grey scale partial mix, 2/3 grey scale mix, One channel partial grey</td>
<td>grey_mix, grey_partial, grey_scale, onechannel</td>
</tr>
<tr>
<td>Denoising</td>
<td>JPEG compression, Gaussian blur, Median filter, Mean filter, Mean bilateral filter, Chambolle denoising, Wavelet denoising, No-local means denoising, Wiener filter</td>
<td>jpeg, bart, gaussian_blur, median_filter, mean_filter, mean_bi_filter, chambolle, wavelet_ran, nonlocal_mean_ran, wiener_filter</td>
</tr>
</tbody>
</table>

Table 2: Strong and Weak Input Transformations for FGSM and Carlini/Wagner on CIFAR-10.

<table>
<thead>
<tr>
<th>Attack</th>
<th>Defense Effectiveness</th>
<th>4 Selected Transformations</th>
</tr>
</thead>
<tbody>
<tr>
<td>FGSM</td>
<td>Strong</td>
<td>wiener_filter, mean_filter, median_filter, chambolle</td>
</tr>
<tr>
<td>Carlini/Wagner</td>
<td>Weak</td>
<td>color_reduction, lab, mean_bi_filter, yuv</td>
</tr>
<tr>
<td>Carlini/Wagner</td>
<td>Strong</td>
<td>chambolle, mean_filter, median_filter, padding</td>
</tr>
<tr>
<td>Carlini/Wagner</td>
<td>Weak</td>
<td>color_reduction, lab, yuv, nonlocal_mean_ran_denoising</td>
</tr>
</tbody>
</table>

3.1 Adversarial Attacks

We chose a variety of adversarial example attacks to generate adversarial examples. They are Fast Gradient Sign Method (FGSM) (untargeted, \( L_\infty \)), Projected Gradient Descent (PGD) (untargeted, \( L_\infty \)), Carlini/Wagner Attacks (targeted, \( L_0, L_2, L_\infty \)), and BPDA, which include not only the common baseline attacks, but also the benchmark attacks such as the BPDA attack used in BaRT.

3.2 A Large Collection of Input Transformation Techniques

Our objective is to assess the effectiveness of BaRT’s approach (Raff et al., 2019), and develop efficient yet robust input transformation ensembles tailored for enterprise users. To achieve this, we have gathered and implemented 33 input transformations, encompassing all techniques employed in BaRT. These transformations were categorized into nine groups, as detailed in Table 1, using (1) reducing the bit depth of each color pixel, (2) introducing noise, (3) rotating pixels, (4) perturbing images, (5) resizing and padding, (6) adding random constant values, (7) enhancing contrast, (8) transforming RGB-colored images to grayscale, and (9) eliminating semantically irrelevant regions in images, to defend against adversarial examples.

3.3 Program Implementation and Experimental Environment

Dataset and Model. We chose to use CIFAR-10 and ImageNet, and pre-train deep neural network (DNN) model architectures, Carlini for CIFAR-10 and InceptionV3 for ImageNet, to evaluate the effectiveness of each input transformation ensemble against adversarial examples. The mean confidence of the two DNN models in predictions on legitimate examples are 77.96% and 76.276% respectively.

Adversarial Example Generation Methods. We leveraged the code from cleverhans, Madry Lab (PGD-Attack, ), Carlini\_nn\_robust\_attacks (RobustML, ) to implement the attack approaches described in 3.1. Specifically, we implemented white box attacks \textit{FGSM} with \( \epsilon = 0.01, 0.05, 0.1 \), \textit{PGD} \( L_\infty \), Carlini/Wagner \( L_0, L_2, L_\infty \) with target = next, and BPDA.

Input Transformation Implementation. We implemented all input transformation methods by using Python and standard imports including \textit{numpy} and \textit{skimage} from Python libraries. All transformation functions take an array of size \( 32 \times 32 \times 3 \) for CIFAR-10 images and size \( 299 \times 299 \times 3 \) for ImageNet images as input.

Detection Method. We adopted the detection method mentioned in (Xu et al., 2017) to evaluate the effectiveness of input transformations and their ensembles.
against the adversarial examples. The key idea is to compare the pre-trained ML model’s prediction on an original input example with its prediction on the transformed input example. If the transformed input produces a substantially different output from the original input, the system classifies the input image as an adversarial example.

**Experiment Environment.** We conducted our experiments on a NVIDIA DGX-1 server featuring 8 P100 GPU accelerators, dual socket Intel Xeon CPUs (512GB DDR4-2133 RAM), and four 100Gb InfiniBand network interface cards.

### 4 COMPUTATIONALLY EFFICIENT BUT STRONG DEFENSE CONSTRUCTION

In this section, we show a thorough analysis of the detection rates obtained from our experiments, and discuss how we utilize the analysis results to construct a computationally efficient defense.

#### 4.1 Initial Observation of the Experiment Results

As the first step of our experiment, we evaluated the detection rates of the following input transformations and their ensembles against both pre-trained models as mentioned in 3.3.

1. Each of the 33 input transformation techniques.
2. The ensembles of any two of the 33 input transformations.
3. The ensembles of any five of the 33 input transformations. Each of them is randomly chosen from five different groups.

We have executed the program on numerous occasions, accumulating a substantial dataset that comprised 48,000 detection rates. Our analysis shows:

1. The same input transformation technique provides different detection rates against different attacks.
2. A small number of transformations, such as adaptive histogram equalization, are stronger than many other transformations, which provide strong defenses against most attacks.
3. Most ensembles of input transformations provide stronger defenses as the number of input transformations increases. However, it is not always true that an ensemble composed of more input transformations (a longer ensemble) certainly provides a stronger defense than ensembles composed of fewer input transformations (a shorter ensemble). Many shorter ensembles provide stronger defenses than some longer ensembles.

4. PGD attack is a stronger attack than BPDA that was used as a strong attack in the BaRT paper. In our testing, where we subjected input transformation ensembles to PGD $L_{\infty}$ attack on a pre-trained Inception V3 model, 9% of them exhibited a detection rate of 0, while 32% had a detection rate below 50%. Notably, certain ensembles with such low detection rates were constructed using five or more transformations (i.e. JPEG compression, FFT Perturbation, YUV, Gaussian, Color Reduction). This underscores that employing an input transformation ensemble, as outlined in BaRT, does not inherently ensure a robust defense.

#### 4.2 A Further Analysis of the Input Transformation Ensembles

We further refined our experiment to delve into constructing a defense that is both computationally efficient and robust.

In this phase, we employed two types of attacks—FGSM ($\epsilon = 0.01, 0.05, 0.1$) and Carlini/Wagner ($L_{\infty}$ with target = next)–on the pre-trained Carlini model for defense analysis. Initially, each of the 33 input transformations underwent 100 iterations against the two attacks. The four strongest and weakest defenses (detailed in Table 2) were selected from the results. Subsequently, we combined these defenses in various ways—pairing strong with strong, weak with weak, strong with weak, three strong, and two strong with one weak—and tested them against the two attacks. Our key findings include: (1) Ensembles outperform individual transformations in detection rates; (2) Two strong ensembles surpass both two weak ensembles and mixed strong-weak ensembles; (3) The order of strong and weak transformations in an ensemble affects the detection rate minimally; (4) Ensembles with three strong transformations exhibit the strongest defenses; (5) Once an ensemble is sufficiently strong, adding another strong or weak transformation yields marginal improvement in detection rates.

Our analysis is presented with sample results in Table 3, 4, 5 and Figure 2 and 3. Table 3 displays detection rates for random combinations of any two of the four strong defenses. Table 4 shows detection rates for random combinations of any two of the four weak defenses. Table 5 illustrates the detection rates for combinations of one weak and one strong defense. By comparing the "Lowest," "Average," and "Highest" rows in these tables, it’s evident...
that the two strong input transformation ensembles in Table 3 offer stronger defenses than those in Table 4 (two weak ensembles) and Table 5 (mixed weak and strong ensembles). Figure 2 depicts sample detection rates for combinations of two strong and one weak input transformation, and Figure 3 shows sample detection rates for combinations of three strong input transformations (to include more results, in addition to using the four strong input transformations in Table 2, we added one more strong input transformation, “rescale”, and used box plots to show the detection rates). Notably, Figure 2 reveals that combinations of two strong and one weak input transformation techniques have higher detection rates than corresponding subsets of two strong ones in Table 3. However, they exhibit weaker defenses than the three strong input transformation ensembles shown in Figure 3.

To construct computationally efficient input transformation ensembles, we assessed the computational cost (i.e., run time) of each input transformation and their ensembles. Table 6 highlights that the computational expense for the same input transformations, when used against attacks in the ImageNet dataset, is significantly higher compared to those in the CIFAR-10 dataset. Additionally, the computational time of an ensemble rises with the number of input transformations. Our analysis indicates the computational time is primarily influenced by image size, the trained model, and the number of the input transformations. Specifically, applying an input transformation technique to an ImageNet image takes longer than applying it to a CIFAR-10 image. Furthermore, a more extensive input transformation ensemble on the same image incurs higher computational costs, given the sequential execution of each transformation function. Importantly, the computational time for each transformation among our 33 collected techniques, applied to the same dataset and pre-trained model, exhibits minimal variation.

### 4.3 Sample Computational Efficient but Strong Defense

Based on the preceding analysis, we draw the conclusion that ensembles consisting of three strong image input transformations effectively balance computational efficiency with robust defense. To verify this
conclusion, we conducted additional experiments, initially testing all three input transformation ensembles derived from the five selected strong transformations (including "rescale" and the other four listed in Table 2) against various FGSM, Carlini/Wagner, and PGD attacks on the pre-trained Carlini model. Among the 70 results, only 5 instances involving three ensembles demonstrated detection rates between 96% and 98%, while all others achieved a 100% detection rate. We replicated these experiments on ImageNet with pre-trained Inception V3 and ResNet-50 models, obtaining similar results. This reaffirms that ensembles comprising three strong input transformations provide robust defenses (nearly 100% detection rate) against state-of-the-art adversarial examples. In contrast to using ensembles of five random input transformations as proposed in the BaRT paper, which may not consistently achieve high detection rates (some falling below 50%, as discussed in section 4.1) and involves longer computational time, our method ensures a strong defense with reduced computational cost and enhanced semantic value.

5 CONCLUSION

In this work, we assess the effectiveness of using input transformation ensembles to defend against state-of-the-art adversarial attacks. To comprehensively evaluate the widely held belief that an extensive barrage of input transformations ensures robust defense, we collected 33 input transformation techniques covering nearly all known methods. We systematically tested these techniques in various ensembles against state-of-the-art attacks—FGSM, PGD, Carlini/Wagner, and BPDA—considered the strongest benchmark attack, on both CIFAR-10 and ImageNet datasets. Our findings reveal two key insights: (1) PGD emerges as the strongest attack among state-of-the-art adversarial examples; (2) a large ensemble, as proposed in BaRT (five transformations), does not guarantee robust defense. Instead, our experiments demonstrate that three strong input transformation ensembles offer a computationally efficient yet strong defense.

REFERENCES


Carlini, N. and Wagner, D. (2017a). Adversarial examples are not easily detected: Bypassing ten detection meth-


