Adversarial Example Defenses: Ensembles of Weak Defenses are not Strong

Warren He
UC Berkeley

James Wei
UC Berkeley

Xinyun Chen
UC Berkeley

Nicholas Carlini
UC Berkeley

Dawn Song
UC Berkeley

Abstract

Ongoing research has proposed several methods to defend neural networks against adversarial examples, many of which researchers have shown to be ineffective. We ask whether a strong defense can be created by combining multiple (possibly weak) defenses. To answer this question, we study three defenses that follow this approach. Two of these are recently proposed defenses that intentionally combine components designed to work well together. A third defense combines three independent defenses. For all the components of these defenses and the combined defenses themselves, we show that an adaptive adversary can create adversarial examples successfully with low distortion. Thus, our work implies that ensemble of weak defenses is not sufficient to provide strong defense against adversarial examples.

1 Introduction

Neural networks have achieved great performance on a wide range of application domains; in particular, they have demonstrated accuracy comparable to or better than humans on datasets in the fields of image recognition [10] and speech recognition [29]. However, recent work has shown that deep learning models are susceptible to *adversarial examples*: inputs that are similar to a correctly classified input, but which are misclassified [27]. Research on other applications of neural networks has also encountered adversarial examples on different tasks beyond image classification, including deep policies in reinforcement learning and generative models [14, 13, 12, 2, 18]. Adversarial examples pose serious threat especially in security-critical autonomous systems such as self-driving cars. Recent work has shown that adversarial examples remain even when subject to the lossy channel of being photographed [16].

Developing effective defenses against adversarial examples is an important topic. Despite many attempts [26, 7, 9, 20, 8, 6, 5, 11], there is no strong defense against adversarial examples to date.

In this paper, we ask the question: *if we ensemble multiple weaker defenses to adversarial examples, will the combined defense be significantly stronger than the original individual defense?* If it did, then an approach to constructing a robust defense to adversarial examples could be to join together many defenses, each of which independently is only slightly effective, but together are strong. To the best of our knowledge, we are the first to investigate this question.

Towards answering this question, we study three instances of ensembled defenses. First, we study two recently proposed defenses, feature squeezing [30] and the *specialists+1* ensemble method [1], each of which ensembles multiple defenses that compensate for each other's weaknesses. Note that feature squeezing and the specialists+1 ensemble are explicitly designed to combine component defenses that work well together, with the intention of creating a stronger defense.

To study the question of ensemble defense in a broader scope, we also evaluate an ensemble of three independent, mutually compatible detection mechanisms [6, 20, 5]. This represents an approach that combines defenses which were not designed to be used together.

The works that introduce these defenses showed that they are effective at detecting attacks generated for the classifier models to which they are applied. However, we find that neither the components of these defenses nor the combined defenses are effective against an attacker that is aware of the defense.

We evaluate these defenses with new attacks, specific to the defenses. Our attacks are able to defeat all aforementioned defenses with low distortion. From this, we conclude that combining weak defenses, even ones that seem to work well together, does not always perform better against an adaptive attacker. Future work may find an effective combination of defenses, but present approaches are not sufficiently strong.

Contributions We make the following contributions:

- We create effective attacks on feature squeezing [30], including individual squeezing methods and the combined adversarial example detection scheme.
- We create an effective attack on the ensemble-of-specialists defense [1].
- We create effective attacks on an ensemble of recently proposed detectors. We show that adversarial examples can bypass an ensemble of detectors with nearly as little distortion as needed for the strongest individual detector.
- Our results show that ensembled defenses do not provide significantly more resilience against adversarial examples than each individual component included in the ensemble. This implies that an ensemble of weak defenses is not sufficient to provide a strong defense against adversarial examples.
- Our evaluation demonstrates that adaptive adversarial examples transfer across several defense or detection proposals. This phenomenon may provide one reason to explain why ensembling is not an effective approach to building defense mechanisms against adversarial examples.

The rest of this paper is organized as follows: in Section 2, we state the problem and provide background; in Section 3, we describe our attacks on individual feature squeezing defense components and their composite defense; in Section 4, we describe our attack on the ensemble-of-specialists defense; in Section 5, we describe our attack on a composite defense that combines multiple independently proposed detection networks; and in Section 6 we summarize our findings.

2 Overview

We start with an overview of the necessary background, and define the threat models we use and the problem statement and setup for our experiments.

2.1 Background: Adversarial Examples

Recent work has pointed out that deep learning models are vulnerable to adversarial examples: these models give incorrect predictions on inputs that are slightly different from correctly predicted ones [27, 7, 22, 25].

Specifically, suppose we have a classifier F with model parameters θ (we may omit θ for brevity when the context is clear). Let x be an input to the classifier with corresponding ground truth label y. An adversarial example x^* is some instance in the input space that is close to x by some distance metric $d(x,x^*)$ which causes F_θ to produce an incorrect output. Here we only consider those x satisfying $F_\theta(x) = y$.

Prior work considers two classes of adversarial examples. First, an *untargeted* adversarial example is an instance x^* that causes the classifier to produce any incorrect output: $F_\theta(x^*) \neq y$. Second, a *targeted* adversarial example is an x^* that causes the classifier to produce a specific incorrect output y^*: $F_\theta(x^*) = y^*$ where $y \neq y^*$.

Many methods for generating adversarial examples compute the gradient of the model in the direction of a misclassification, with respect to the input image. Several approaches have been proposed in previous work. The Fast Gradient Sign Method (FGSM) [7] and Fast Gradient Method [19] take a fixed-size step in the direction of a misclassification, with FGSM using the sign of the direction. These generate images at a fixed L_∞ distance and L_2 distance from the original image, respectively (modulo image box constraints). The Jacobian-based Saliency Map Approach (JSMA) [25] greedily modifies pixels that, based on the gradient, would most strongly reduce the confidence of a correct classification. This generates images at a fixed L_0 distance from the original image. DeepFool [21] extends the fast gradient methods with an iterative approach, similar to gradient descent optimization. Optimization-based methods [27, 4, 19] take this approach further with more sophisticated optimization algorithms and specialized loss functions.

2.2 Threat Models

In this work, we assume the adversary has full knowledge of the model, including the model architecture, parameters, and the defense strategies used in the model. Prior work has shown this assumption can often be relaxed. Adversarial examples generated for one model can successfully fool other models, even ones models of different architectures and models trained on different data [7, 23]. An attacker can thus train a model of its own and generate adversarial examples to fool a black-box model [24, 19]. For simplicity we assume this stronger white-box threat model.

Within these white-box attackers, we consider two capacities of adversaries.

Static Adversary. A *static adversary* is an attacker that is not aware of any defenses that may be in place to protect the model against adversarial examples. A static adversary can generate adversarial examples using existing methods but does not tailor attacks to any specific defense.

Adaptive Adversary. An *adaptive adversary* is an attacker that is aware of the defense methods used in the model, and can adapt attacks accordingly. This is a strictly more powerful adversary than static adversary. In this paper, we consider this stronger adversary.

2.3 Problem Statement

To improve the robustness of models against adversarial examples, prior work investigates into two directions. The first direction attempts to produce correct predictions on adversarial examples, while not compromising the accuracy on legitimate inputs [26, 7, 9]. The other (more recent) direction instead attempts to *detect* adversarial examples, without introducing too many false positives. In this case, the model can reject an instance and refuse to classify those that it detects as adversarial [20, 8, 30, 1].

As stated earlier, in this paper, we ask the question: *if we ensemble multiple defenses to adversarial examples, then will the combined defense be significantly stronger than each individual original defense?* This is an important question for designing effective defense against adversarial examples, as if the answer were "yes," then it may be possible to construct a strong defense simply by ensembling multiple weaker defenses. In this paper we argue the answer is "no": by considering many defenses, we show that ensembled defenses are not always more robust.

Defenses considered. In this paper, we consider defenses that attempt to combine multiple (somewhat weaker) defenses to construct a larger strong defense. In particular, we look at three instances of ensemble defense strategies. First and second are feature squeezing [30] and the *specialists+1* ensemble method [1], both of which take this approach by construction. These defenses are constructed from components that are intended to be useful together. Their authors have shown that these defenses effectively detect low-perturbation adversarial examples generated by a static adversary. Third, to study the effectiveness of ensembling defenses more broadly, we merge together many detectors that were not designed to be used in conjunction with any other detector. In particular, as an example demonstration, we ensemble three independent detection mechanisms [6, 20, 5] to build one detection mechanism.

For each of these defense strategies, we propose attack methods to generate adversarial examples as an adaptive adversary against the individual component defense (when applicable) as well as the composite defense strategy. We use these attack methods to evaluate each component defense and composite defense: if our method succeeds at generating adversarial examples, this means that an adaptive adversary can defeat the defense. To gauge how strong the combined defense is compared to the components, we compare the level of distortion needed to fool each (using the same optimization method).

2.4 Experimental Setup

Datasets and models. To evaluate the effectiveness of the different defense strategies, we use two standard datasets, MNIST [17] and CIFAR-10 [15] datasets.

For both datasets, we randomly sample 100 images in the test set, filter out examples that are not correctly classified, and generate adversarial examples based on the correctly classified images. When evaluating each defense strategy, we use the same model architectures described in their papers respectively [30, 1, 6, 20, 5].

Our experiments took up to three minutes to generate each adversarial example. The attacks we use can scale up to larger models, which require more computation per optimization step. On the other hand, prior work has shown that larger models are actually easier to fool, with lower-distortion adversarial examples or better success at a fixed level of distortion [7, 21, 28, 4]. Our own results agree, with adversarial examples on a ResNet for CIFAR-10 having significantly lower distortion than adversarial examples on a smaller CNN for MNIST (a much smaller dataset). We expect even larger datasets would be even easier to attack.

Adversarial example generation method. In this paper, we use an optimization-based approach to generate adversarial examples [4], which is shown to be effective on finding adversarial examples with small distortions.

At a high level, the attack uses an optimizer to minimize a loss function:

$$\text{loss}(x') = \|x' - x\|_2^2 + c \cdot J(F_\theta(x'), y)$$

Here, F_θ is a part of the trained classifier that outputs a vector of logits, and J computes some penalty based on the logits and some label y, either a ground truth label for non-targeted attacks or a target label for targeted attacks. A constant c is a hyperparameter that adjusts the relative weighting between distortion and misclassification. We omit details of the design choice and refer the reader to the original paper [4].

Measurement of distortion. Unless otherwise specified, we measure the distortion between an adversarial example and the original input as the L_2-norm of their distance. Formally, $d(x^*, x) = \sqrt{\sum_i (x_i^* - x_i)^2}$. Each dimension of input images is scaled to $[0, 1]$, i.e., $0 \leq x_i^*, x_i \leq 1$.

3 Adaptive attacks on feature squeezing

In this and next section, we investigate ensemble defense strategies that are intentionally constructed to have component defenses which work together to detect adversarial examples. The first defense we study is *feature squeezing*, proposed by Xu et al. [30, 31].

Background: feature squeezing defense. To perform feature squeezing, one generates a lower fidelity version of the input image through a process known as "squeezing" before passing it into the classifier. Xu et al. proposed two methods of squeezing: reducing the color depth to fewer bits, and spatially smoothing the pixels with a median filter. According to their paper, the two methods of squeezing work well together because they address two major kinds of perturbation used in adversarial examples: color depth reduction eliminates small changes to many pixels, while spatial smoothing eliminates large changes to a few pixels.

In order to detect adversarial examples, Xu et al. propose a system combining the two squeezing methods. First, the system runs the classifier on three different versions of the image: the original image, the reduced-color-depth version and the spatially smoothed version of the original image. Then, it compares the softmax probability vectors across these three classifier outputs. The L_1 *score* of the input is the highest L_1 distance between any pair of softmax probability vectors among the three. It flags inputs where the L_1 score exceeds a threshold as adversarial.

In their experiments, Xu et al. show that MNIST and CIFAR-10 classifiers are accurate on squeezed inputs. On adversarial examples generated by a static adversary using FGSM [7] and JSMA [25], they show that their detector achieves 99.74% accuracy on a test set with equal portions benign and adversarial examples. They also show that squeezing the input alone prevents 84 – 100% of the adversarial examples (correctly classifying them). Recently, Xu et al. showed that a simplified detector that uses the original version of the input and the spatially-smoothed version (excluding the color-depth-reduced version) achieves a 98.80% overall detection accuracy on MNIST and 87.50% on CIFAR-10 against a static adversary using a variety of Carlini and Wagner's attacks [31].

Summary of our approach and results. We demonstrate that feature squeezing is not an effective defense in two stages. First, we show that an adaptive attacker can construct an adversarial example that remains adversarial after it is squeezed by each method (color depth reduction and spatial smoothing. Then, we use this approach to construct adversarial examples that are classified the same way both with and without squeezing, causing the L_1 score to be smaller than a given fixed threshold. Our results show that the combined detection method is not effective against an adaptive attacker.

3.1 Evading individual feature squeezing defense components

In these experiments, we evaluate whether adversarial examples are robust to each individual feature squeezing defense component, i.e., whether adversarial examples remain adversarial after each individual feature squeezing process (color depth reduction and spatial smoothing) separately. These experiments attack the components of the combined feature squeezing detection scheme. Performing this attack is necessary for defeating the combined detection scheme, wherein the predicted label probabilities of squeezed images are compared against each other.

3.1.1 Evading color-depth-reduction defense

The first method of squeezing an image that Xu et al. propose is color depth reduction. This method rounds each value in the input to 2^b evenly spaced values spanning the same range, which we refer to as reducing to b bits.

Attack Approach. We use the method described in Section 2.4 to generate adversarial examples that are robust to color depth reduction. After each step of the optimization procedure, an intermediate image (perturbed from the original image) is available from the optimizer. We check if a reduced-color-depth version of this intermediate image is adversarial. We run the optimization multiple times, initializing the optimization with random perturbations of the original image each time, so that it explores different optimization paths. For each original image, we keep the successful adversarial example that has the lowest L_2 distance to the original image among all the generated successful adversarial examples for this original image.

Attack results on MNIST. We evaluate color depth reduction to 1 – 7 bits. On the strongest defense evaluated by Xu et al., which reduces color depth to 1 bit, we successfully generated adversarial examples for *all* original images, with an average distortion of 3.86. Figure 1 shows a sample of these adversarial examples. Table 1 summarizes our results for other bit depths. Notice that for a system without any color depth reduction (retaining the original 8 bits of depth), we find adversarial examples with an average distortion of 1.38. Reducing color depth to fewer bits makes the system less sensitive

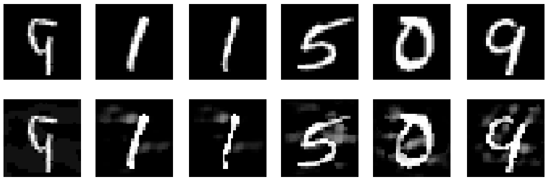

Figure 1: Adversarial examples for color depth reduction (to 1 bit) on MNIST. First row: original images. Second row: adversarially perturbed. Distortions, from left to right: 1.49, 2.61, 2.63, 3.83, 3.89, 3.90.

Bit depth	Adv success	Avg L_2
1	100%	3.86
2	99%	1.69
3	100%	1.43
4	100%	1.39
5	100%	1.44
6	100%	1.33
7	100%	1.33
8	100%	1.38

Table 1: Summary of MNIST adversarial examples that are misclassified when reduced to different color depths. “Adv success” measures the fraction of original images for which we successfully found an adversarial example. “Avg L_2” measures the average L_2 distortion of the successful adversarial examples.

to small changes, which requires larger distortions; however, the distortions are still very small.

Attack results on CIFAR-10. We evaluate color depth reduction to 3 bits, which Xu et al. recommend as a good balance between the accuracy on adversarial inputs and accuracy on benign images for CIFAR-10. We succeeded at generating adversarial examples for *all* original images, with an average distortion of 0.945. Figure 2 shows a sample of these adversarial examples. For comparison, adversarial examples for a classifier without color depth reduction have an average distortion of 0.214. Although this method of squeezing increases the distortion needed for successfully generating non-targeted adversarial examples using the same optimization method, again, such a distortion is still small and imperceptible.

Summary. An adaptive attacker can successfully generate adversarial examples with small distortions for a system that applies color depth reduction to the input image before classifying it.

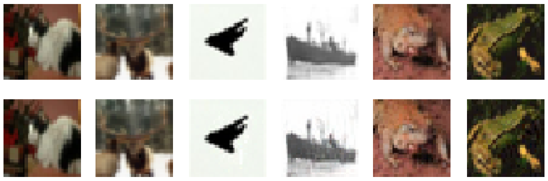

Figure 2: Adversarial examples for color depth reduction (to 3 bits) on CIFAR-10. Distortions, from left to right: 0.0194, 0.0954, 0.322, 0.942, 0.948, 0.948. Layout is the same as Figure 1.

Filter size	Adv success	Avg L_2
3×3	100%	1.29
2×2	100%	1.57
5×5	100%	0.612
3×1	100%	1.33
1×3	100%	1.29
2×1	100%	1.52
1×2	100%	1.51
5×1	100%	0.943
1×5	100%	0.931

Table 2: Summary of MNIST adversarial examples that are misclassified when spatially smoothed with varying sizes of median filters. Columns have the same meaning as in Table 1. Some filters make adversarial examples *easier* to find.

3.1.2 Evading spatial smoothing

Xu et al. propose a second method for feature squeezing, which applies a median filter to the input, which replaces each pixel with the median value of a neighborhood around the pixel.

To generate adversarial examples that are misclassified after spatial smoothing, we use the procedure from Section 2.4 with the addition of a median filter as part of the classification model. A median filter for TensorFlow was not available, so we implemented our own.

Attack results on MNIST. We evaluate a range of median filter sizes, ranging from 1×2 to 5×5. For a 3×3 filter, with which Xu et al. achieved the best accuracy, we successfully generated adversarial examples for *all* original images, with an average distortion of 1.29. Figure 3 shows a sample of these adversarial examples. Table 2 summarizes our results for other filter sizes. Larger median filters did not require greater distortion. Compared to adversarial examples generated for a system without any spatial smoothing (average distortion of 1.38), the average distortion is not increased.

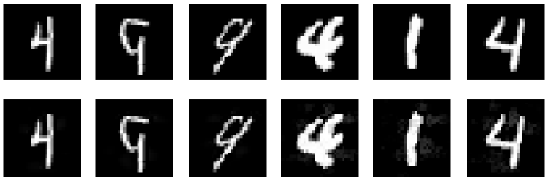

Figure 3: Adversarial examples for spatial smoothing (with 3 × 3 filter) on MNIST. Distortions, from left to right: 0.236, 0.241, 0.282, 1.27, 1,31, 1.31. Layout is the same as Figure 1.

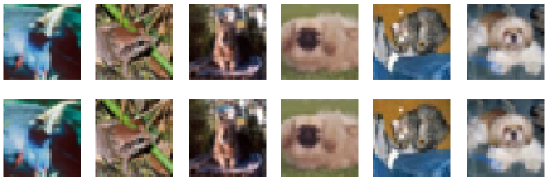

Figure 4: Adversarial examples for spatial smoothing (with 2 × 2 filter) on CIFAR-10. Distortions, from left to right: 0.0273, 0.0537, 0.0584, 0.198, 0.211, 0.212. Layout is the same as Figure 1.

Attack results on CIFAR-10 We evaluate a 2 × 2 median filter, which Xu et al. identify as achieving a good rejection rate of adversarial examples and accuracy on benign images on CIFAR-10. We successfully generated adversarial examples for *all* original images, which have an average distortion of 0.205. Figure 4 shows a sample of these adversarial examples. The average distortion is not higher than for a system without spatial smoothing (0.214).

Summary Spatial smoothing alone is not an effective defense against an adaptive attacker. We have shown that an adaptive adversary can create adversarial examples for a system that applies spatial smoothing which are not more distorted than adversarial examples for a baseline system that does not apply spatial smoothing.

3.2 Evading combination of multiple squeezing techniques

While the individual feature squeezing techniques are weak against an adaptive attacker, we investigate whether the detection scheme that combines them is stronger. In this case, we find that this detection scheme is not much stronger than the strongest component defense, color depth reduction.

Background: Composite feature squeezing defense The detection scheme combines both methods of squeezing. In particular, the detection system has three "branches," where each one runs the classifier on a different version of the input, the original input, a reduced-color-depth version and a spatially-smoothed version of the original input. These three branches output different softmax probability vectors, and the scheme compares the L_1 distance between these vectors against a threshold to determine whether the input is adversarial.

In this experiment, we evaluate the effectiveness of the combined defense against an adaptive attacker, i.e., whether adversarial examples can be misclassified and bypass this detection scheme.

Attack Approach. We use the procedure from Section 3.1.1 with modifications to support the detection mechanism. In this attack, we run two instances of the model in parallel during optimization, representing the detector branches that operate on the input and the spatially smoothed version of the input. The optimizer receives gradients from both models, equally. We do not include the reduced-color-depth branch in the gradient calculation, because the color depth reduction removes gradient information; it is, however, included when we compute the L_1 score. We collect only adversarial examples that have an L_1 score below a threshold of 0.3076, a level at which Xu et al. achieved the best accuracy in their experiments on MNIST.

Attack results on MNIST We evaluate a combination of color depth reduction to 1 bit and smoothing with a 2 × 2 median filter, which Xu et al. found to be accurate on adversarial examples generated by a static adversary [31]. We successfully generated adversarial examples for *all* original images, with an average distortion of 4.76 and L_1 score of 0.209. Figure 5 shows a sample of these adversarial examples. These examples are misclassified and successfully evade detection. This distortion is 23.3% larger than for color depth reduction alone, but still very small.

Attack results on CIFAR-10. We evaluate a combination of color depth reduction to 3 bits and smoothing with a 2 × 2 median filter, a combination of settings that perform well in Xu et al.'s experiments. We successfully generated adversarial examples for *all* original images, with an average distortion of 0.601 and L_1 score of 0.168. Figure 6 shows a sample of these adversarial examples. These examples are misclassified and successfully evade detection. This distortion is even lower than that of the color depth reduction defense alone. Although Xu et al. do not prescribe a threshold specific to CIFAR-10, the

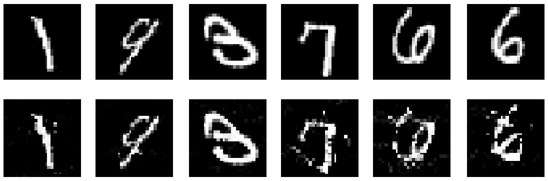

Figure 5: Adversarial examples for combined feature squeezing detection on MNIST. Distortions, from left to right: 2.00, 2.04, 2.39, 4.66, 4.77, 4.79. Layout is the same as Figure 1.

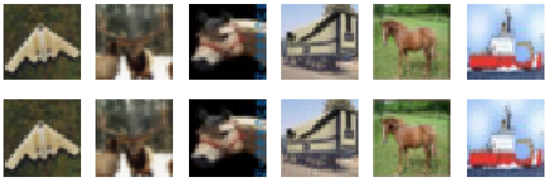

Figure 6: Adversarial examples for combined feature squeezing detection on CIFAR-10. Distortions, from left to right: 0.117, 0.120, 0.130, 0.604, 0.614, 0.617. Layout is the same as Figure 1.

average L_1 score for these examples is lower (i.e., detected as less adversarial) than the average L_1 score for the original images, which is 0.225.

Summary. The detection scheme that combines two methods of squeezing is not always stronger than the strongest component, color depth reduction. The improvement is low even on MNIST, which is particularly well suited for feature squeezing, with images being black and white (little change from color depth reduction) and having large, contiguous areas of the same color (little change from spatial smoothing). On CIFAR-10, the combined attack requires less distortion than the color depth reduction defense alone.

4 Evading ensemble of specialists

We study a second defense that combines multiple component defenses, an ensemble of specialists, proposed by Abbasi and Gagné [1].

Background: ensemble of specialist defense. The defense consists of a generalist classifier (which classifies among all classes) and a collection of specialists (which classify among subsets of the classes). The specialists classify subsets of the classes as follows. Where C is the set of all K classes in the task, for each class i, let U_i be the set of classes with which i is most often confused in adversarial examples. To compute U_i, Abbasi and Gagné select the top 80% of misclassifications caused by nontargeted FGSM attacks for each class i. Further, K additional subsets are defined: $U_{K+i} = C \setminus U_i$ to be the complement set of U_i. For each $j = 1, \ldots, 2K$, a specialist classifier F_j is trained on a subset of the dataset containing images belonging to the classes in U_j to classify input images into the classes in U_j only. In addition, a generalist classifier F_{2K+1} is trained to classify input images into classes in $U_{2K+1} = C$. Each classifier in the ensemble may be susceptible to basic adversarial examples, but the proposed defense assumes that each specialist can detect a few specific attacks, thus the attacker cannot fool all specialists and the generalist at the same time. The defense combines them to jointly detect general adversarial examples.

In order to classify an input, the system first checks if, for any class i, the generalist classifier and all specialists that can classify i agree that the input belongs to class i. If such a class i exists, note that at most one class can get the generalist's vote, it must be unique. In this case, the system takes the mean of the outputs of the generalist and the specialists that can classify i. Otherwise, at least one classifier has misclassified the input, and the system takes the mean of the outputs of all classifiers in the ensemble.

Abbasi and Gagné [1] find that using an ensemble constructed this way successfully reduces the system's confidence (mean confidence among classifiers activated by the voting scheme) on adversarial examples generated by a static attacker using FGSM [7], DeepFool [21], and Szegedy et al.'s approach [27]. They conclude that a classification system can use an ensemble of diverse specialists this way and detect low-confidence examples as adversarial.

Attack approach. In this experiment, we evaluate the effectiveness of Abbasi and Gagné's specialists+1 ensemble against an adaptive attacker. We considered a scenario where a user provides an image to a system, and

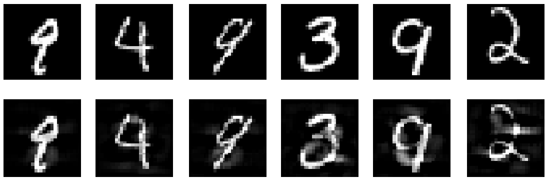

Figure 7: Adversarial examples for specialists+1 on MNIST. Distortions, from left to right: 1.55, 1.76, 1.83, 3.77, 3.90, 3.93. Layout is the same as Figure 1.

the system uses a specialists+1 ensemble to classify the image or reject it as adversarial.

We attempt to create targeted adversarial examples, where we chose target classes randomly. For each original image, then our goal is to create an adversarial example that is classified as the target class by the generalist classifier and all applicable specialists at the same time, and with high confidence from those classifiers. We use the procedure from Section 2.4 to generate adversarial examples. In this experiment, we kept only adversarial examples that were misclassified with confidence greater than the average confidence on a sample of benign images, 0.999708. We modified the loss function to support multiple classifiers:

$$\mathrm{loss}(x') = \|x' - x\|_2^2 + c \sum\nolimits_{j\in\{1,\ldots,2K+1\}; y^*\in U_j} J(F_j(x'), y^*)$$

We evaluate this defense on MNIST only. While Abbasi and Gagné also propose the defense for CIFAR-10, the architecture described in their experiments have low accuracy on CIFAR-10, resulting in low confidence even in benign images.

Attack results on MNIST. We successfully generated adversarial examples for *all* original images, which have an average distortion of 3.87. Figure 7 shows a sample of these adversarial examples in the second row. These adversarial examples are classified as the target label by the generalist and all applicable specialists. For comparison, the average confidence of a single generalist classifier on correctly classified benign images is 0.998951, and a batch of targeted adversarial examples with at least that confidence has average distortion 3.65. The distortion needed for high-confidence adversarial examples on specialists+1 is 6.03% higher than for a non-ensemble MNIST classifier.

Although this defense defines the specialists to focus on common misclassifications caused by non-targeted adversarial examples, it is still weaker at detecting the common misclassifications. Among the examples, 33 targeted a class that the original image's ground truth class was commonly confused with. The average distortion for these images is 3.06, below the average of the entire set.

Summary. The specialists+1 ensemble does not effectively ensure low confidence on adversarial examples generated by an adaptive attacker. An adaptive attacker can successfully generate adversarial examples with small distortions, which are unanimously classified as a target class, and thus evade the detection of the specialist+1 ensemble defense.

		Source Defense		
		Gong	Metzen	Feinman
Target	Gong	100%	51%	21%
	Metzen	43%	100%	18%
	Feinman	96%	92%	100%

Table 3: Probability that adversarial examples constructed for a given source defense also fool the given target defense on CIFAR-10. Defenses generated against Metzen *et al.* transfer to the others with the highest probability, and Feinman *et al.* with lowest.

5 Evading ensemble of detectors

In the previous sections, we have investigated ensembles of defenses that are intentionally constructed to be useful together. In Xu et al.'s work, the color depth reduction is intended to remove small changes to many pixels, and the median smoothing to remove large changes to a few pixels. Similarly, Abbasi and Gagné propose using an ensemble of generalist and specialist classifiers together; without the others, this approach would not be useful.

To study the effectiveness of ensembling defenses more broadly, we merge together three recently proposed detectors that were not designed to be used in conjunction with any other detector. We consider only detectors that are applied to a fixed classification network for simplicity, and therefore study the following schemes:

- Gong et al. propose using adversarial training to detect adversarial examples [6]. Given the original model, generate adversarial examples on the training data. Then, train a new classifier that distinguishes the original training data from the adversarial data.
- Metzen et al. construct a similar scheme, however instead of using the original images as the input to the detector, they train on the inner convolutional layers of the network [20].
- Feinman et al. examine the final hidden layer of a neural network and find that adversarial examples are separable from the original images by training a density estimate using Gaussian kernels [5].

When using Carlini and Wagner's attack, these approaches are known to provide only slight increases in robustness, i.e., only increase the required distortion slightly when generating the adversarial examples with the detector vs. without the detector [3]. Given this, we now examine if constructing an ensemble of these defenses provides additional robustness. To ensemble these defenses, we run each detection method and report the input as adversarial if any of the three detectors do.

Attack approach. We perform this experiment on CIFAR-10 exclusively, as Metzen et al.'s defense is intended for a ResNet applied to CIFAR-10. We are able to construct adversarial examples for all defenses independently. To defeat all three defenses together, we construct a new classifier $G(\cdot)$ so that using the loss function from Section 2.4 directly can construct adversarial examples.

We use the same notation as [3]. Let $F(\cdot)$ be a classifier on N classes, and $\text{softmax}(F(\cdot))_i$ be the probability of class i (so that $F(\cdot)_i$ are the logits). Let $\{D_j(x)\}_{j=1}^{J}$ be one of J different detectors so that the probability that detector D_j reports object x as adversarial is $\text{sigmoid}(D_j(x))$ (that is, D_j returns the logits). We report that an instance is an adversarial example if the probability of any detector is greater than one half. That is, if for any j, $\text{sigmoid}(D_j(x)) > \frac{1}{2}$, or, alternatively, $D_j(x) > 0$.

When we ensemble the three defenses, we set $J = 3$ and define $D(x) = \max_j D_j(x)$, so that $D(x)$ reports adversarial (i.e., $D(x) > 0$) if any of the three detectors do.

Given this, we use the same $G(\cdot)$ construction as Carlini and Wagner's previous work on these defenses [3]. This function $G(\cdot)$ returns N+1 classes (with the new class reserved for adversarial examples) so that $\arg\max_i G(x)_i = \arg\max_i F(x)_i$ when x is not adversarial, and $\arg\max_i G(x) = N+1$ when x is adversarial. To do this, [3] specifically defines

$$G(x)_i = \begin{cases} F(x)_i & \text{if } i \leq N \\ (D(x)+1)\cdot \max\limits_j F(x)_j & \text{if } i = N+1 \end{cases}$$

If for a given instance x, $D_j(x) > 0$ (for any classifier j) then we will have $\arg\max_i G(x)_i = N+1$ since we multiply a value greater than one by the largest of the other output logits. Conversely, if $\arg\max_i G(x)_i \neq N+1$ then we must have $D(x) < 0$ implying that all detectors report the instance is benign.

Therefore, by constructing adversarial examples on G so that the target class is not $N+1$, we can construct adversarial examples on F that are not detected by any detector.

Attack results on CIFAR-10 The L_2 distortion required to construct adversarial examples on an unsecured network is 0.11. To construct adversarial examples on this network $G(\cdot)$ with the three defenses increases the distortion to 0.18, an increase of 60%. However, this distortion is still imperceptible.

Transferability of adversarial examples across different detectors. In order to understand the reason that these defenses do not significantly increase robustness when combined together, we hypothesize that the transferability property [27, 7, 23, 19] of adversarial examples is simplifying the attacker's task. To verify this, we construct adversarial examples on each of the three defenses in isolation and check the probability that these examples also fool the other two defenses. Table 3 contains this data. Feinman's defense is the weakest of the three, and so transfers least often (and adversarial examples transfer to it most often). The other two defenses are approximately equally effective. From this, we can see one possible reason why constructing an ensemble of these weak defenses is not significantly more secure than each independently: the adversarial examples that fool one detector may also fool the other detectors. We conclude that one must be careful when ensembling defenses to build them to cover the weaknesses of the others, and not simply assemble them blindly.

6 Conclusion

In this paper we examine whether multiple (possibly weak) defenses can be combined to create a strong defense. Towards this goal, we studied three such defenses that combined multiple components: two recently proposed defenses designed with rationale of why their components should work well together and one that combined unrelated recently proposed detectors.

We showed that an adaptive adversary can generate adversarial examples with low distortion that fool all of the defenses that we evaluate. The feature squeezing detection scheme, which combines two methods of squeezing an input image, is at best marginally stronger than color depth reduction alone. The specialists+1 ensemble, which combines several specialist classifiers, increases the required distortion slightly, but again, distortion is still small. We also showed that combining a collection of recently proposed detection mechanisms is also ineffective. In particular, our results show that adversarial examples transfer across the individual detectors.

Our work sheds light on a few important lessons when evaluating defenses against adversarial examples: 1) one should evaluate defenses using strong attacks. For example, FGSM can quickly generate adversarial examples, but may fail to generate successful attacks when other iterative optimization based methods can succeed; 2) one should evaluate defenses using adaptive adversaries. It is important to develop defenses that are secure against attackers who know the defense mechanisms being used.

Finally, our results indicate that combining weak defenses does not automatically improve the robustness of these systems. Developing effective defenses against adversarial examples is an important topic. We hope our work sheds light for future work in this area.

Acknowledgements We thank Weilin Xu, David Evans, and Mahdieh Abbasi for discussing their work with us. We thank Arjun Bhagoji, Chang Liu, and Richard Shin, who have provided valuable feedback. This work was supported in part by BDD (Berkeley Deep Drive); the CLTC (Center for Long-Term Cybersecurity); FORCES (Foundations Of Resilient CybEr-Physical Systems), which receives support from the National Science Foundation (NSF award numbers CNS-1238959, CNS-1238962, CNS-1239054, CNS-1239166); and the National Science Foundation under Grant No. TWC-1409915. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Cloud computing resources were provided through a Microsoft Azure for Research award.

References

[1] Abbasi, M., and Gagné, C. Robustness to adversarial examples through an ensemble of specialists. *5th International Conference on Learning Representations (ICLR) Workshop* (2017).

[2] Behzadan, V., and Munir, A. Vulnerability of deep reinforcement learning to policy induction attacks. *13th International Conference on Machine Learning and Data Mining (MLDM)* (2017).

[3] Carlini, N., and Wagner, D. Adversarial examples are not easily detected: Bypassing ten detection methods. *arXiv preprint arXiv:1705.07263* (2017).

[4] Carlini, N., and Wagner, D. Towards evaluating the robustness of neural networks. *38th IEEE Symposium on Security and Privacy (SP)* (2017).

[5] Feinman, R., Curtin, R. R., Shintre, S., and Gardner, A. B. Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410* (2017).

[6] Gong, Z., Wang, W., and Ku, W.-S. Adversarial and clean data are not twins. *arXiv preprint arXiv:1704.04960* (2017).

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

[8] Grosse, K., Manoharan, P., Papernot, N., Backes, M., and McDaniel, P. On the (statistical) detection of adversarial examples. *arXiv preprint arXiv:1702.06280* (2017).

[9] Gu, S., and Rigazio, L. Towards deep neural network architectures robust to adversarial examples. *NIPS 2014 Deep Learning and Representation Learning Workshop* (2014).

[10] He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision* (2015), pp. 1026–1034.

[11] Hendrycks, D., and Gimpel, K. Early methods for detecting adversarial images. *5th International Conference on Learning Representations (ICLR) Workshop* (2017).

[12] Huang, S., Papernot, N., Goodfellow, I., Duan, Y., and Abbeel, P. Adversarial attacks on neural network policies. *5th International Conference on Learning Representations (ICLR) Workshop* (2017).

[13] Kos, J., Fischer, I., and Song, D. Adversarial examples for generative models. *arXiv preprint arXiv:1702.06832* (2017).

[14] Kos, J., and Song, D. Delving into adversarial attacks on deep policies. *5th International Conference on Learning Representations (ICLR) Workshop* (2017).

[15] Krizhevsky, A. Learning multiple layers of features from tiny images. Tech. rep., 2009.

[16] Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial examples in the physical world. *5th International Conference on Learning Representations (ICLR) Workshop* (2017).

[17] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE 86*, 11 (1998), 2278–2324.

[18] Lin, Y.-C., Hong, Z.-W., Liao, Y.-H., Shih, M.-L., Liu, M.-Y., and Sun, M. Tactics of adversarial attack on deep reinforcement learning agents. *26th International Joint Conference on Artificial Intelligence* (2017).

[19] Liu, Y., Chen, X., Liu, C., and Song, D. Delving into transferable adversarial examples and black-box attacks. *5th International Conference on Learning Representations (ICLR)* (2017).

[20] Metzen, J. H., Genewein, T., Fischer, V., and Bischoff, B. On detecting adversarial perturbations. *5th International Conference on Learning Representations (ICLR)* (2017).

[21] Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), IEEE, pp. 2574–2582.

[22] Nguyen, A., Yosinski, J., and Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), IEEE, pp. 427–436.

[23] Papernot, N., McDaniel, P., and Goodfellow, I. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277* (2016).

[24] Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., and Swami, A. Practical black-box attacks against deep learning systems using adversarial examples. *ACM Asia Conference on Computer and Communications Security (ASIACCS)* (2017).

[25] Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., and Swami, A. The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)* (2016), IEEE, pp. 372–387.

[26] Papernot, N., McDaniel, P., Wu, X., Jha, S., and Swami, A. Distillation as a defense to adversarial perturbations against deep neural networks. In *Security and Privacy (SP), 2016 IEEE Symposium on* (2016), IEEE, pp. 582–597.

[27] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *International Conference on Learning Representations (ICLR)* (2014).

[28] Tabacof, P., and Valle, E. Exploring the space of adversarial images. In *Neural Networks (IJCNN), 2016 International Joint Conference on* (2016), IEEE, pp. 426–433.

[29] Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., and Zweig, G. Achieving human parity in conversational speech recognition. *IEEE Workshop on Spoken Language Technology* (2016).

[30] Xu, W., Evans, D., and Qi, Y. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155* (2017).

[31] Xu, W., Evans, D., and Qi, Y. Feature squeezing mitigates and detects Carlini/Wagner adversarial examples. *arXiv preprint arXiv:1705.10686* (2017).