The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Towards Imperceptible and Robust Adversarial Example Attacks against Neural Networks

Bo Luo, Yannan Liu, Lingxiao Wei, Qiang Xu
Department of Computer Science & Engineering
The Chinese University of Hong Kong
{boluo,ynliu,lxwei,qxu}@cse.cuhk.edu.hk

Abstract

Machine learning systems based on deep neural networks, being able to produce state-of-the-art results on various perception tasks, have gained mainstream adoption in many applications. However, they are shown to be vulnerable to adversarial example attack, which generates malicious output by adding slight perturbations to the input. Previous adversarial example crafting methods, however, use simple metrics to evaluate the distances between the original examples and the adversarial ones, which could be easily detected by human eyes. In addition, these attacks are often not robust due to the inevitable noises and deviation in the physical world. In this work, we present a new adversarial example attack crafting method, which takes the human perceptual system into consideration and maximizes the noise tolerance of the crafted adversarial example. Experimental results demonstrate the efficacy of the proposed technique.

Introduction

With the increasing popularity of deep learning technology, its security problems have attracted a significant amount of attention from both academia and industry. Among all security problems, the most severe one is adversarial example attack first proposed in (Szegedy et al. 2013). It attempts to modify the legal input with slight perturbations that largely changes the output given by neural networks. This kind of attack is really a threat for security-sensitive systems, such as self-driving systems, disease diagnose systems and malicious email filters (Bojarski et al. 2016; Amato et al. 2013; Clark, Koprinska, and Poon 2003; Graves, Mohamed, and Hinton 2013). For example, in self-driving systems, an adversarial example attack can change a stop-road sign to a turn-left signal. Then the car will make a wrong decision and cause a serious traffic accident.

Recently, several adversarial example attacks against image classification systems are proposed in the literature (Szegedy et al. 2013; Papernot et al. 2016; Goodfellow, Shlens, and Szegedy 2014). They target to misclassify the input image to specific class or just misclassify it by adding minimum perturbations. However, the perturbations imposed on input images by these attacks are usually perceptible and not robust. On the one hand, existing attacks do not consider human perceptual system during perturbation generation. They all use distance metrics of L_p norms (L_0, L_2 and L_∞ norms) to evaluate the similarity between the original samples and the crafted adversarial ones. These metrics treat perturbations of different pixels in an image equally important for human eyes. However, according to (Liu et al. 2010), people are more sensitive to perturbations of pixels in low variance regions. For example, perturbations in the uniform background are easier to be detected than those in image regions with mussy objects. Without considering the human perceptual system, previous methods may perturb highly sensitive pixels in the attack process, which increases the probability to be detected by human eyes. On the other hand, previous attacks are not robust enough. The success rate of attack drops significantly in the physical world due to the noises and deviation inevitably generated. For instance, adversarial images may be compressed or suffer from noises during transmission from the attacker to the classifier. Thus, the adversarial example which attacks successfully in the experimental condition may fail in the complex physical world. Recently, some research efforts have been dedicated to robust attacks for certain situations, such as face recognition (Sharif et al. 2016) and road sign recognition (Evtimov et al. 2017). However, they are rather application-specific and cannot be generalized for other applications.

To solve the above problems, in this paper, we propose a new method to craft imperceptible and robust adversarial examples against neural networks. We first introduce a new distance metric considering sensitivity of the human perceptual system to different pixels. This metric guides us with how many perturbations can be added without being detected. Then we optimize to maximize the noise tolerance of adversarial examples to improve the success attack rate in the physical world, which is generally applicable for a large amount of applications based on neural networks. By introducing a new metric to evaluate the effects of perturbations added to each pixel, we present a greedy algorithm to find which pixels to perturb and what magnitude to add effectively and efficiently that is applicable for most models as long as they are differentiable. Our optimization method can generate adversarial examples with both high imperceptibility and high robustness, as demonstrated in the experimental

results.

Related Work

Szegedy *et al.* first proposed adversarial example attack against neural networks. It minimizes the distances evaluated with L_2-norm between the original examples and the adversarial ones under the constraint that the adversarial attack is successful. FGSM (Goodfellow, Shlens, and Szegedy 2014) performs the attack by first calculating the gradients of the loss function to search which directions to change for each pixels. Then it modifies all pixels simultaneously under the L_∞ constraint. Recently, JSMA (Papernot et al. 2016) builds a saliency map to model the impact each pixel has on the resulting classification. It then optimizes with the L_0 distance, where it picks the most important pixel based on the saliency map and modify the pixel to increase the target class probability in each iteration. However, these attack methods all use simple distance metrics (L_p-norms) to evaluate the similarity between the adversarial example and the original one without considering the human perceptual system.

There are some research efforts about robust adversarial example attacks in the literature. The authors in (Kurakin, Goodfellow, and Bengio 2016) first discussed the idea when they found some adversarial examples survived in the physical world. However, they did not present a solution to improve the success attack rate. Recently, two papers studied adversarial examples for certain applications in the physical world. (Sharif et al. 2016) proposed a physical realizable adversarial example attack against the face recognition system through wearing malicious eye-glasses. (Evtimov et al. 2017) discussed a practical attack on the road sign recognition in self-driving systems. They generate adversarial road signs which can successfully deceive the recognizer in various directions and angles. However, these two methods are rather application-specific and are not generally applicable.

Apart from adversarial example attacks against neural networks in computer vision systems, many other machine learning applications are suffering from adversarial example attacks. (Carlini et al. 2016) proposed an attack against the speech recognition system, where they show how to craft sounds that are difficult for human to understand, but can be interpreted to specific commands such as "Call 911" and "Turn on airplane mode". In (Grosse et al. 2016), they introduce adversarial example attacks against malware detecting systems. In these attacks, they disguise a malware into a benign one and successfully fool the detector.

Adversarial Example Attacks

Adversarial example attacks target to change the output of machine learning systems by adding slight perturbations to the input. In the literature, there are two categories of adversarial example attacks: target attack and un-target attack. For the target attack, it attempts to misclassify a sample to a specific class, while un-target attack only tries to misclassify the input. As a result, the target attack is more difficult than the un-target one. In this paper, we focus on the target adversarial example attack.

Generally speaking, adversarial example attacks should not only fool machine learning systems but also consider two important factors: imperceptibility and robustness.

Imperceptibility: The imperceptibility of an adversarial example means that it should look similar to the original one in order not to be detected by human eyes. So in the attack, it is important to use an appropriate distance metric to evaluate the visual similarity between an adversarial example and the original one. A good distance metric should clearly reflect the characteristic of the human perceptual system.

Robustness: Adversarial examples are firstly crafted by the attackers and then transmitted to the machine learning systems. They may fail to attack after the transmissions with inevitable noises or deviation. The robustness of adversarial examples reflects its ability to stay misclassified to the target class after the transformations in the physical world. The definition is as follows:

$$\begin{aligned} F(X^*) &= T, \\ F(Tran(X^*)) &= T, \end{aligned} \tag{1}$$

where X^* is the adversarial example crafted by the attacker, T is the target class specified, and $Tran(*)$ is the transformation in the physical world. Previous methods do not consider the robustness and thus adversarial examples crafted by them may be largely destroyed and fail to attack in the physical world.

In this paper, we propose a new crafting method to generate adversarial examples with both high imperceptibility and high robustness, as detailed in the following sections.

The Proposed Method

In this section, we first present a new distance metric considering the effects of different pixels on human eyes, then we propose to maximize the gap between the probability of target class and the max probability of left classes to increase the noise tolerance of adversarial examples. Lastly, an efficient greedy algorithm is introduced which can craft adversarial examples with high imperceptibility and high robustness.

Imperceptibility of Adversarial Example Attacks

According to contrast masking theory in image processing (Legge and Foley 1980; Lin, Dong, and Xue 2005; Liu et al. 2010), human eyes are more sensitive to perturbations on pixels in low variance regions than those in high variance regions. For example, in Figure 1, the left image is the original sample. The middle and right images are perturbed with the same magnitude on ten pixels but at different positions. The positions of perturbations on the middle image are all at high variance region (the mussy desk) while the right image is perturbed at the low variance region (the black bag on the floor). We can see that it is hardly to detect the perturbations on the middle image, however, people with normal visual capability can notice the perturbations on the black bag.

Therefore, to make adversarial examples imperceptible, we should perturb pixels at high variance zones rather than

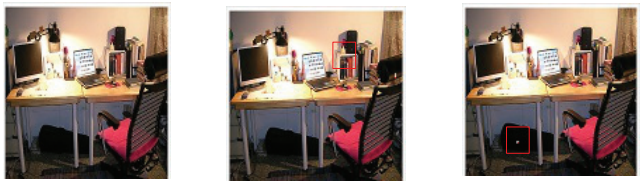

Figure 1: Perturbations added in different pixels raise varying human perceptual attention. The red line box marks the perturbation in each perturbed image.

low variance ones. In this paper, we compute the variance of a pixel x_i based on the standard deviation $SD(x_i)$ among an $n \times n$ neighborhood of pixel x_i as shown in Equation 2, where S_i is the set consisting of pixels in the $n \times n$ region, μ is the average value of pixels in the region. Specifically, for $n = 3$, the variance is calculated as the standard deviation of the pixel and its 8 neighbors.

$$SD(x_i) = \sqrt{\frac{\sum_{x_k \in S_i} (x_k - \mu)^2}{n^2}}. \tag{2}$$

Accordingly, we introduce *perturbation sensitivity* to measure how much "attention" will be drawn by adding per "unit" perturbation on a pixel. It is defined as follows:

$$Sen(x_i) = 1/SD(x_i). \tag{3}$$

When the pixel is in a low variance region, the perturbation sensitivity is high. Therefore, adding perturbations on this pixel is easily detected by human eyes.

To evaluate the human perceptual effect of a perturbation added to a pixel, we can multiply the magnitude of the perturbation by its sensitivity. When crafting an adversarial example, we usually perturb more than one pixel. As a result, we sum up all the effects of perturbations and use it as the distance between the original example and the adversarial one, as shown in the following equation:

$$D(X^*, X) = \sum_{i=1}^{N} \delta_i * Sen(x_i), \tag{4}$$

where $D(X^*, X)$ denotes the distance between the adversarial example X^* and the original one X. δ_i is the perturbation added to the pixel x_i and N is the total number of pixels.

Robustness of Adversarial Example Attacks

Another limitation of existing methods for adversarial example attack is that they have very low success rates in the physical world due to deviation caused by regular transformations of images such as compressing, resizing and smoothing. The challenge of the problem is that transformations in the physical world are usually uncertain and hard to model, and thus we cannot enhance robustness of attacks for specific situations. In this paper, we give a general solution for robust attacks by maximizing noise tolerance of adversarial examples. The noise tolerance reflects the amount of noises that adversarial examples can tolerate with the misclassified target label unchanged.

Neural-network-based classifiers output the probabilities for all classes, and select the highest one as the result label for the given input. The probability for one class denotes the confidence of classifying the input to this category. Previous adversarial example attacks only maximize the probability of the target class, however, we find for robust attack, it is necessary to reduce the probability of left classes as well. Naturally, we dedicate to maximize the gap between the probability of the target class and the maximal probability of all other classes. It can be formulated as follows:

$$Gap(X^*) = P_t - max(P_i) \qquad (i \neq t), \tag{5}$$

where P_t denotes the target class probability and P_i refers to probabilities of other classes. Intuitively, the higher the probability gap, the more robust the adversarial example attacks.

Figure 2 is a simple example to illustrate this idea. In this figure, there are two adversarial examples against the same original sample. They are all misclassified as a ship with 0.6 probability, but with different probability gap. Adversarial example 1 has a higher probability gap than adversarial example 2. Now, after JPEG compression (quality is 60), adversarial example 1 still is classified as ship with 0.5 probability while adversarial example 2 is classified as dog with 0.52 probability and the probability of ship now decreases to 0.36. Only the first two classes with the highest probability are listed in the figure.

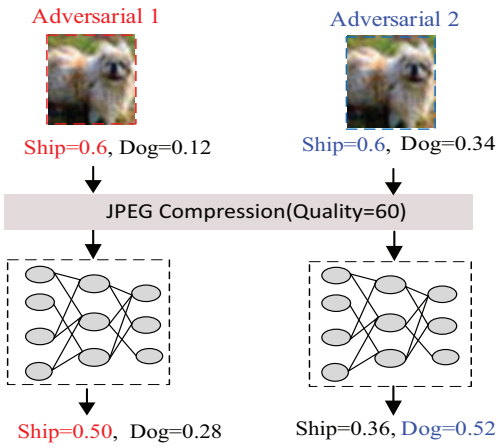

Figure 2: Adversarial example with higher probability gap is more robust when suffering the same image transformation.

Imperceptible and Robust Attacks

In this work, we target to achieve both imperceptible and robust adversarial example attacks against neural networks. In order to obtain imperceptibility, the distance $D(X^*, X)$ between the original example X and the adversarial one X^* should be constrained. Then we can increase the robustness of adversarial example attacks whenever possible under this constraint. Overall, we formulate the problem as follows:

$$\begin{aligned} &\underset{X^*}{\text{argmax}} \;\; Gap(X^*) \\ &s.t. \;\; D(X^*, X) \leq D_{max} \end{aligned} \tag{6}$$

where D_{max} is the largest distance allowed in order not to be detected by human eyes. In practice, users need to determine this value based on the input images. There are two difficulties in solving this optimization problem. Firstly, the objection function is not differentiable as it contains a max function so that gradient descent, a common technique to solve optimization problem, is not applicable. Secondly, optimization problems are usually solved in an iterative manner and in each iteration, we should determine which pixels to perturb and how many perturbations to add. The search space in each iteration is tremendously large so that an efficient algorithm is desired.

To deal with these problems, we first smooth the objective function to make it differentiable and then propose an efficient greedy algorithm to simplify computations in each iteration, as detailed in the following subsections.

Smoothing the Objective Function We smooth the max function to the differentiable one based on the following equation:

$$max(x, y) \approx \log(e^{kx} + e^{ky})/k. \tag{7}$$

It achieves the approximation by amplifying the difference between kx and ky using the exponential function. When kx is quite bigger than ky, e^{kx} will be much larger than e^{ky}. Then $e^{kx} + e^{ky}$ will approximately equal to e^{kx} and $log(e^{kx})/k$ is essentially equal to x. When x and y are not significantly different, k is used to improve the accuracy of approximation. Given an example with $x = 0.2, y = 0.1$, if $k = 1$, then $\log(e^{kx} + e^{ky})/k \approx 0.84$. However, it approximates to 0.2000005 when $k = 100$. We can make the approximation as close to the max function as we want by setting large enough k.

Now the objective function is transformed in the following format and can be differentiated for further optimization:

$$Gap(X^*) \approx P_t - \log(\sum e^{kP_i})/k \qquad i \neq t \tag{8}$$

A Greedy Algorithm for Optimization After smoothing the objective function, we can solve the problem using the traditional gradient descent method. However, in each iteration, we have to choose which pixels to modify and what magnitudes to add. Even though we assume each pixel is perturbed with the same magnitude, the time taken for solving the problem is still prohibitively long. For example, if each image contains 100 pixels and we choose to perturb 10 pixels at each iteration, then we have to search $\binom{100}{10}$ times to find the best 10 pixels to modify.

Considering that we have to choose pixels with less perturbation sensitivity to human eyes and at the same time increase the objective function in Equation 8, we define a new metric called *perturbation priority* to estimate the effect of perturbing a pixel:

$$PerturbPriority(x_i) = \frac{\bigtriangledown_{x_i} Gap(X^*)}{Sen(x_i)}, \tag{9}$$

where $\bigtriangledown_{x_i} Gap(X^*)$ is the gradient of the probability gap for pixel x_i. Perturbation priority indicates how much probability gap increased by adding one "unit" of perturbation to the current pixel x_i, and therefore it reflects the priority of pixels to perturb in the adversarial example generating process.

Algorithm 1: The proposed algorithm to generate adversarial examples.

Input: The legitimate sample X, the max allowed human perceptual distance D_{max}, the number of pixels perturbed in each iteration m and the perturbation magnitude δ.
Output: Adversarial example X^*.

1 **while** $D(X, X^*) < D_{max}$ **do**
2 $\quad PerturbPriority \leftarrow$ Calculate perturbation priority for each pixel;
3 $\quad SortedPerturbPriority \leftarrow$ Sort perturbation priority in $PerturbPriority$;
4 $\quad SelectedPixels \leftarrow$ Choose m pixels with largest perturbation priority;
5 $\quad X^* \leftarrow$ Perturb selected pixels with magnitude δ;
6 $\quad D(X^*, X) = \sum_i^N \Delta_i * Sen(x_i)$;
7 **end**

Based on the perturbation priority, we propose a greedy algorithm to efficiently achieve imperceptible and robust adversarial example attacks. The detailed crafting process is shown in Algorithm 1, in which it first calculates each pixels' perturbation priority based on the gradients of the probability gap and perturbation sensitivity in line 2. Then we sort pixels according to perturbation priority in line 3. Next, we perturb the first m pixels with a small magnitude δ and calculate the human perceptual distance of the updated adversarial examples in line 4-6, where Δ_i is the total perturbations added to x_i. The whole process is repeated until the constraint on $D(X^*, X)$ is violated.

Experimental Evaluations

Dataset. All the experiments are performed on MNIST and CIFAR10 datasets. The MNIST dataset (LeCun, Cortes, and Burges 2010) includes 70000 gray scale hand-written digit images with the size of 28*28. The classification goal is to map the images to the corresponding digits from 0 to 9. The CIFAR10 dataset (Krizhevsky, Nair, and Hinton 2014) contains 6000 color images. Each image has the size of 32*32*3. There are 10 classes in the dataset, which are airplane, automobile, bird, cat, deer, dog, frog, horse, ship and truck. The intensity values of pixels in all these images are scaled to a real number in $[0, 1]$.

DNN Model. For each dataset, we trained a model. The architectures of these two models are commonly used in corresponding classification tasks detailed in Table 1. They are all 8 layers DNN with ReLU as the activation function. The MNIST and CIFAR10 model achieve 99.18% and 84.21% classification rate respectively.

Baselines. The baselines used in these experiments are three widely-used adversarial example attacks, Jacobian-based Saliency Map Approach (JSMA) (Papernot et al. 2016), iterative Fast Gradient Sign Method (FGSM) (Goodfellow, Shlens, and Szegedy 2014) and box-constrained L-

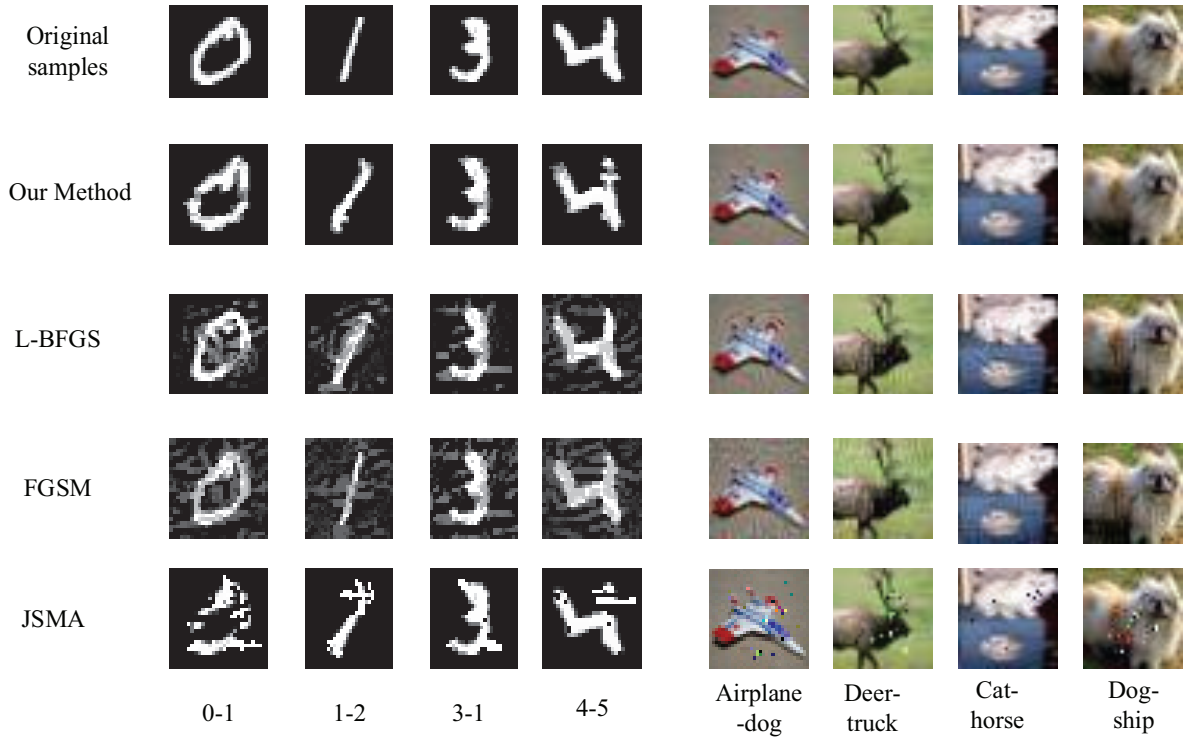

Figure 3: Adversarial examples generated by different crafting methods in MNIST and CIFAR10. Adversarial examples in the second row crafted by our method are much more imperceptible than others from the following rows. While JSMA method in the last row performs the worst.

Table 1: Model architectures.

Layer	MNIST	CIFAR
Input layer	28, 28	32, 32, 3
Convolution layer 1	3, 3, 32	3, 3, 64
Convolution layer 2	3, 3, 32	3, 3, 64
Max pooling layer 1	2, 2	2, 2
Convolution layer 3	3, 3, 64	3, 3, 128
Convolution layer 4	3, 3, 64	3, 3, 128
Max pooling layer 2	2, 2	2, 2
Fully connected layer 1	128	512
Fully connected layer 2	10	10
Softmax		

BFGS method (Szegedy et al. 2013). For detailed experimental setups of these methods please refer to the original papers. In our method, we select 20 pixels to add perturbations with a magnitude of 0.01 in each iteration.

Evaluate Imperceptibility

In this experiment, we evaluate the imperceptibility of adversarial examples crafted by our method and the baseline methods. We perform adversarial example attacks against the testing set (10000 test images) in MNIST and CIFAR10 respectively. These adversarial examples are just successfully misclassified to the target classes which were randomly assigned. That is to say, we stop adding perturbations once the target class attack is successful. Then for each attack method, we get two groups of adversarial examples for the two datasets.

Evaluate with Human Perception: We randomly choose several groups of images to present in Figure 3. Each group images include an original sample and its corresponding adversarial examples crafted by different attack methods. The left four columns are from MNIST and the right four columns are from CIFAR10. Images in the first row are original samples in the testing set. The second row are adversarial examples crafted by our method. The following rows are adversarial examples from L-BFGS, FGSM and JSMA attack methods, respectively. The target adversarial classes are listed at the bottom.

From Figure 3, we can see that adversarial examples crafted by our method in the second row are the most imperceptible, which is nearly the same as the original one. While JSMA method in the last row performs the worst and the perturbed pixels are easily detected by human eyes. The reason is that JSMA method perturbs pixels to the maximum value without considering pixels' human perceptual sensitivity. As a result, the perturbed pixels may be in the high sensitive region and thus raise human attentions. For L-BFGS and FGSM methods, they perform better than JSMA. This is because these two methods use L_2 norm and L_∞ norm which tend to perturb more pixels with smaller perturbations. Although the pixels perturbed may be in the sensi-

tive region, they raise relatively low attentions with smaller perturbations. These experimental results show that considering human perceptual system, our method can generate much more imperceptible adversarial examples comparing with baseline methods.

Evaluate Distance Metric: To evaluate the effectiveness of our human perceptual distance metric, we calculate the distances between the adversarial examples and original samples. The results are listed in Figure 5.

We can see that the distances of adversarial examples crafted by our method are the smallest (44.78 for MNIST and 51.98 for CIFAR10), while the distances of JSMA method are the largest (80.34 for MNIST and 92.25 for CIFAR10). This coincides with the results in the previous human perception experiments. Therefore, we can believe the distance metric proposed in this work can appropriately reflect the visual similarity between original samples and adversarial examples.

Discussions: From the above results, we know that human perceptual distances in MNIST are larger than those in CIFAR10, so adversarial example attacks against MNIST dataset are more difficult than CIFAR10 dataset. We analyze the reasons from two aspects. One is that the images in MNIST have a large uniform background, while for CIFAR10, the backgrounds of natural images only occupy small regions. As a result, in MNIST images, pixels have a higher human perceptual sensitivity than those in CIFAR10. The other reason is that the classification rate in MNIST dataset is about 15% higher than CIFAR10. So the model trained in MNIST has higher confidence with the classifying results, thus it is more difficult to attack the model trained with MNIST dataset.

Evaluate Robustness

In this experiment, we evaluate the robustness of adversarial examples crafted with our method and the baseline methods. We compare all the attack methods under the same max human perceptual distance, $D_{max} = 70$. This is determined empirically that the distance less than 70 would not raise much human attention. In this part we only present the results for dataset CIFAR10, because the results for MNIST are quite similar.

Robustness Definition: The robustness can be described as the fraction of adversarial examples which are still misclassified as the target class after the natural transformations. It is also called the success attack rate in the physical world. The definition is as follows:

$$R = \frac{\sum_{i=1}^{m} C(X_i, label_i) C(X_i^*, T_i) C(Tran(X_i^*), T_i)}{\sum_{i=1}^{m} C(X_i, label_i) C(X_i^*, T_i)} \quad (10)$$

where m is the number of testing samples used to compute the robustness. X_i is a test image and $label_i$ is the true label of this image, and X_i^* is the adversarial example. T_i is the target class for X_i sample assigned by the attacker. The function $Tran(*)$ is an image transformation operation in the physical world. We study several transformations in this experiment, including adding gaussian noises, JPEG compressing, image blurring, changing contrast and brightness. The function $C(X, label)$ is used to check whether the image was classified correctly or not:

$$C(X, label) = \begin{cases} 1, & \text{If image } X \text{ is classified as } label; \\ 0, & \text{otherwise.} \end{cases}$$

Evaluate Robustness with Gaussian Noises and Transformations: We test the physical success rate using four image transformations: JPEG compressing, gaussian blurring, contrast and brightness adjusting. The experimental results are showed in Figure 4. We also test the robustness with gaussian noises which have five intensities with standard deviation changed from 0.05 to 0.25 with a step of 0.05 (The gaussian mean is 0). The experimental results are listed in Table 2.

Table 2: Comparison of robustness for various adversarial methods adding gaussian noises.

Noises	Our	JSMA	L-BFGS	FGSM
Std=0.05	98.5%	98.25%	86.8%	82.5%
Std=0.1	94.0%	88.5%	82.0%	79.5%
Std=0.15	77.8%	68.8%	62.6%	64%
Std=0.2	68.5%	55.12%	50.8%	42.5%
Std=0.25	62%	33.2%	28.6%	21.5%

It is clearly shown that our method performs the best among all the four transformations. For example, in JPEG compressing, our method performs 76% success attack rate while for FGSM method, the success rate is only 52.3%. Experimental results in adding gaussian noises show that our method also achieves higher robustness than other ones. Moreover, the benefit is more obvious with stronger noises. For example, in the fifth intensity with standard deviation (0.25), our method achieves 62% success rate while the average success rate of the baseline methods is just about 26%.

Apart from these observations, there are other observations drawn based on the results. Firstly, we can see that JSMA method performs the second best in these experiments though it achieves the worst results in previous human perception experiments. It can be explained that JSMA method perturbs less pixels with larger perturbations, and these large perturbations can tolerant more noises added on them. While for the FGSM method, it has the worst robustness in these experiments, because it tends to make small perturbations on the whole image. The effects of these small perturbations on pixels are more easily changed with noises. Our method, however, tries to maximize the noise tolerance and at the same time consider imperceptibility, therefore, it achieves great results for both imperceptibility and robustness.

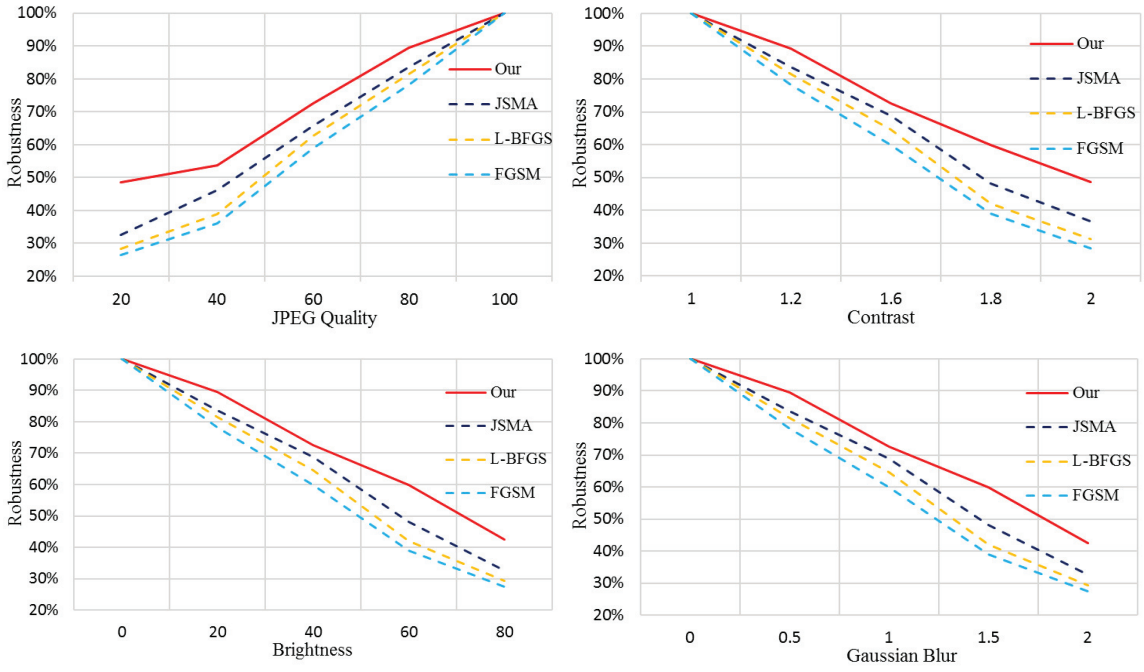

Figure 4: Comparison of robustness for various adversarial methods for image transformations.

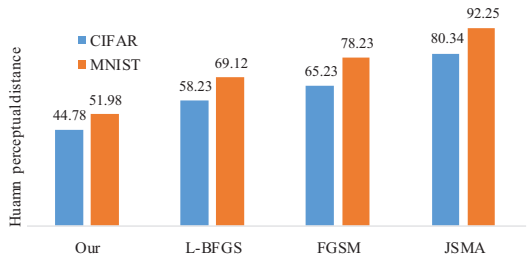

Figure 5: Human perceptual distances of adversarial exampels crafted by varying adversarial methods in MNIST and CIFAR10. Adversarial examples crafted by our method has the smallest human perceptual distances.

Conclusions

Adversarial example attacks against neural networks have become one of the most severe security problems in artificial intelligence. Traditional adversarial example attacks do not consider human perceptual systems and thus are easily detected. Moreover, the success attack rate drops largely due to inevitable noises in the physical world. In this paper, we introduce a new adversarial example attack, which can achieve both high imperceptibility and robustness in the physical world. Specifically, we propose a new distance metric considering human perceptual systems. The metric evaluates the sensitivity of image pixel to human eyes, and thus it can guide us to add perturbations with less chances of being detected. To improve the successful attack rate in practice, we try to maximize the probability gap between the adversarial target class and other classes. A simple yet effective greedy algorithm is introduced to achieve the optimization goal under the constraint of not being detected. Experimental results show that adversarial examples crafted by our method is more imperceptible and robust than those produced by previous methods.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China (NSFC) under Grant No.61432017 and 61532017, and in part by a research grant provided by InbesTech Co. Ltd.

References

Amato, F.; López, A.; Peña-Méndez, E. M.; Vaňhara, P.; Hampl, A.; and Havel, J. 2013. Artificial neural networks in medical diagnosis.

Bojarski, M.; Del Testa, D.; Dworakowski, D.; Firner, B.; Flepp, B.; Goyal, P.; Jackel, L. D.; Monfort, M.; Muller, U.; Zhang, J.; et al. 2016. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*.

Carlini, N.; Mishra, P.; Vaidya, T.; Zhang, Y.; Sherr, M.; Shields, C.; Wagner, D.; and Zhou, W. 2016. Hidden voice commands. In *USENIX Security Symposium*, 513–530.

Clark, J.; Koprinska, I.; and Poon, J. 2003. A neural network based approach to automated e-mail classification. In *Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on*, 702–705. IEEE.

Evtimov, I.; Eykholt, K.; Fernandes, E.; Kohno, T.; Li, B.; Prakash, A.; Rahmati, A.; and Song, D. 2017. Robust physical-world attacks on machine learning models. *arXiv preprint arXiv:1707.08945*.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Graves, A.; Mohamed, A.-r.; and Hinton, G. 2013. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, 6645–6649. IEEE.

Grosse, K.; Papernot, N.; Manoharan, P.; Backes, M.; and McDaniel, P. 2016. Adversarial perturbations against deep neural networks for malware classification. *arXiv preprint arXiv:1606.04435*.

Krizhevsky, A.; Nair, V.; and Hinton, G. 2014. The cifar-10 dataset. *online: http://www. cs. toronto. edu/kriz/cifar. html*.

Kurakin, A.; Goodfellow, I.; and Bengio, S. 2016. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*.

LeCun, Y.; Cortes, C.; and Burges, C. J. 2010. Mnist handwritten digit database. *AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist* 2.

Legge, G. E., and Foley, J. M. 1980. Contrast masking in human vision. *Josa* 70(12):1458–1471.

Lin, W.; Dong, L.; and Xue, P. 2005. Visual distortion gauge based on discrimination of noticeable contrast changes. *IEEE Transactions on Circuits and Systems for Video Technology* 15(7):900–909.

Liu, A.; Lin, W.; Paul, M.; Deng, C.; and Zhang, F. 2010. Just noticeable difference for images with decomposition model for separating edge and textured regions. *IEEE Transactions on Circuits and Systems for Video Technology* 20(11):1648–1652.

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

Sharif, M.; Bhagavatula, S.; Bauer, L.; and Reiter, M. K. 2016. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 1528–1540. ACM.

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