EXPLOITING THE INHERENT LIMITATION OF L_0 ADVERSARIAL EXAMPLES

Fei Zuo, Bokai Yang, Xiaopeng Li, Lannan Luo, Qiang Zeng

22nd International Symposium on
Research in Attacks, Intrusions and Defenses

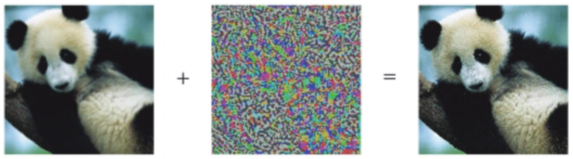
ADVERSARIAL EXAMPLES AND COUNTERMEASURES

USC South Carolina

First we will introduce some necessary background.

NEURAL NETWORKS ARE VULNERABLE TO AE ATTACKS

- Adversarial examples (AEs) are crafted by adding human-imperceptible perturbations to inputs in order that a neural-network-based classifier incorrectly labels them.



Original Image "Giant Panda" | Adversarial Noise | Adversarial Image "Indri"

South Carolina

For example, VGG16 can correctly classify the left image as giant panda.
By contrast, after introducing some subtle noises, the adversarial image can fool the neural networks.

ADVERSARIAL EXAMPLES TAXONOMY

- To quantitatively describe such subtle perturbations, L_{p} norms are usually used to measure the discrepancy between an original benign image and its corresponding AE.
- According to the value of p, the mainstream AE generation algorithms can be categorized into three families, namely L_∞, L_2 and L_0 attacks.
- Informally, L_∞ measures the largest modification among the pixels, L_2 measures the Euclidean distance between the two images, and L_0 measures the number of modified pixels.

This work focuses on L_0 AEs.

Univ of SC. South Carolina

JSMA and CW-L0 are two leading L0 AE generation methods, we consider them both in our paper.

COUNTERMEASURES AGAINST AES

- To defeat attacks based on AEs, both detection and defensive techniques attract the research community's attention.
- Given an input image, the ***detection*** system outputs whether it is an AE, so that the target neural network can reject those adversarial inputs.
- A ***defense*** technique, given an AE, helps the target neural network make correct prediction by either rectifying the AE or fortifying the classifier itself.

This work involves both defense and detection.

CHALLENGES FROM L0 AES

- Many AE detection methods and defense techniques have been proposed. However, prior methods either are not very effective in handling L_0 AEs or omit discussing them.
- Previous work even argues that it is challenging to recover the correct classification of L_0 AEs by input transformation, as "it is very difficult to properly reduce the effect of the heavy perturbation". [Liang et al., 2018]
- For example, bit depth reduction is effective to defense L_2 attacks [Xu et al, 2018]. However, this approach only has a very limited capability to defend against L_0 attacks.

CHALLENGES FROM L0 AES

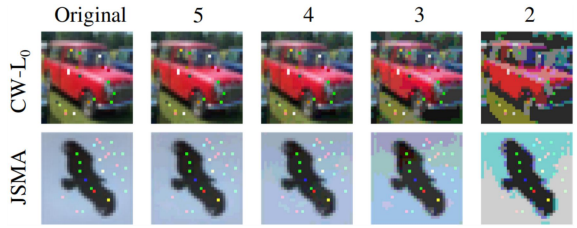


Table: The classification accuracy for AEs after applying bit depth reduction

Dataset	Attack	Bit depth			
		2-bit	3-bit	4-bit	5-bit
CIFAR-10	JSMA	22.2%	27.1%	21.2%	12.0%
	CW-L_0	51.2%	56.6%	55.1%	51.5%

SC South Carolina

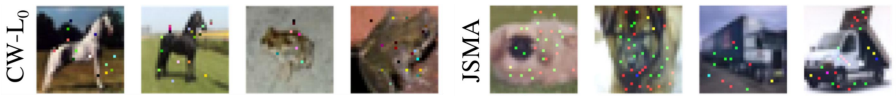
We show some image examples from CIFAR-10 after applying bit depth reduction.
Given the different numbers of bit depth, the first row displays a benign image and its processed versions;
the first row displays an AE generated by CW-L0 and its corresponding processed images;
the second row displays an AE generated by JSMA and its corresponding processed images.

As shown in the Table, processing the AEs generated by JSMA and CW-L0 with bit depth reduction cannot Significantly improve the classification accuracy of the target model.

HOW TO EFFECTIVELY DETECT AND DEFENSE L0 AES?

THE INHERENT CHARACTERISTICS OF L0 AES

- We identify two characteristics of L_0 AEs:
  - The first characteristic is that it limits the number of modified pixels, but not the amplitude of pixels. Thus, L_0 attacks tend to introduce large-amplitude perturbations.
  - Second, as L_0 attacks try to modify as few pixels as possible, the optimization-based AE generation process tends to result in altered pixels that scatter in the image.



South Carolina

In other words, those corrupted parts are mostly small and isolated regions.

Here, we show some concrete adversarial samples generated by CW and JSMA algorithm.

By exploiting the two characteristics, we build the defense and detection system based on a heuristic method and simple architecture to effectively thwart such kind of AE attacks.

DEFENSE STRATEGY

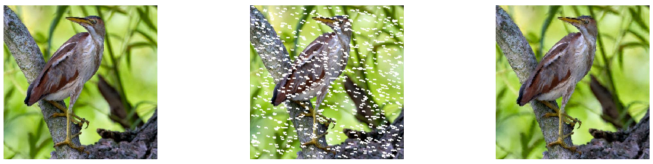
- The aforementioned characteristics of L_0 attacks imply that for an altered pixel, it is highly possible that one extreme value can be observed in at least one channel.
- For example, an original pixel is represented as an intensity vector [0.32, 0.56, 0.62], where all the values are normalized.
- After corrupting by the L_0 attack, it becomes [0.33, 0.55, 0.96], whose B channel has an extreme value 0.96.

U of SC. South Carolina

We define a value as extreme if it is either smaller than an upper bound or larger than a lower bound.
We present more empirical analysis about the range of extreme value. You can refer to our paper for more details.
Here, we show some concrete cases.
The leftmost image is an adversarial example genereted by JSMA algorithm.
The following images are three masks which locate the pixels whose have extreme values in R, G, B channels, respectively.

INPAINTING

- In digital image processing, *inpainting* is the process of reconstructing lost or deteriorated parts of images and videos.



[Telea, A., 2004]

[Elad, M. et al., 2005]

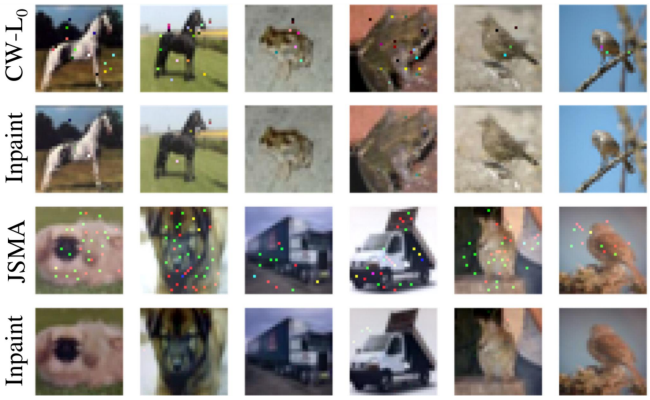
Original image | Corrupted image | Restored image

U of SC. South Carolina

If we can locate those the most likely adversarial pixels based on our heuristic, then we could use inpainting technique to restore these images.

We show some examples here. The leftmost images are original images. Numerous parts are lost in the two corrupted images. After using inpainting technique, they can be well restored and visually recognisable.

DEFENSE EFFECTIVENESS



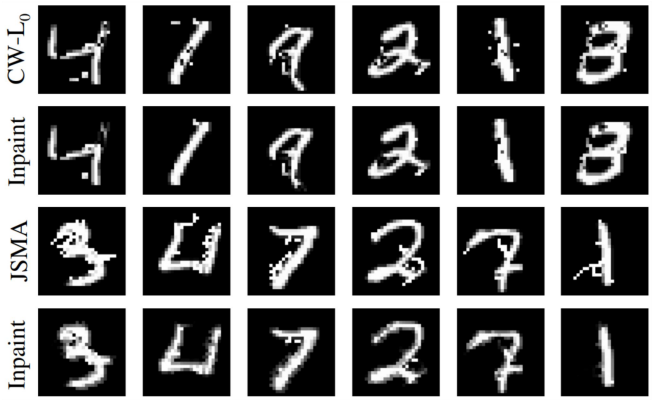
- Dataset: CIFAR-10
- For CW-L_0 and JSMA attack, after using the proposed defense method, the classification accuracy on the AEs is increased from 0% to 87.3%, and from 0% to 96.1%, respectively.

South Carolina

Based on this straight forward strategy, we design a pre-processor to rectify the AEs.
Please refer to our paper for more details of the proposed algorithm.
Here we show some concrete examples.
The first and third rows show the CW-L0 and JSMA attack applied to CIFAR-10 images, respectively.
The second and fourth rows show the corresponding resulting images after restoring.

One important insight is that the masks are unnecessary to be very accurate.
In other words, in an advasarial image, even though one benign pixel is labeled as adversarial by mistake,
the inpainting works very well for recovering it in a benign way.
However, for an adversarial pixel, the inpainting effect usually is not what
the AE attacker desires, since the maliciously perturbed pixels can hardly be recovered to the attacker-intended values.

DEFENSE EFFECTIVENESS



- Dataset: MNIST
- For CW-L_0 and JSMA attack, after using the proposed defense method, the classification accuracy on the AEs is increased from 0% to 88.2%, and from 0% to 86.1%, respectively.

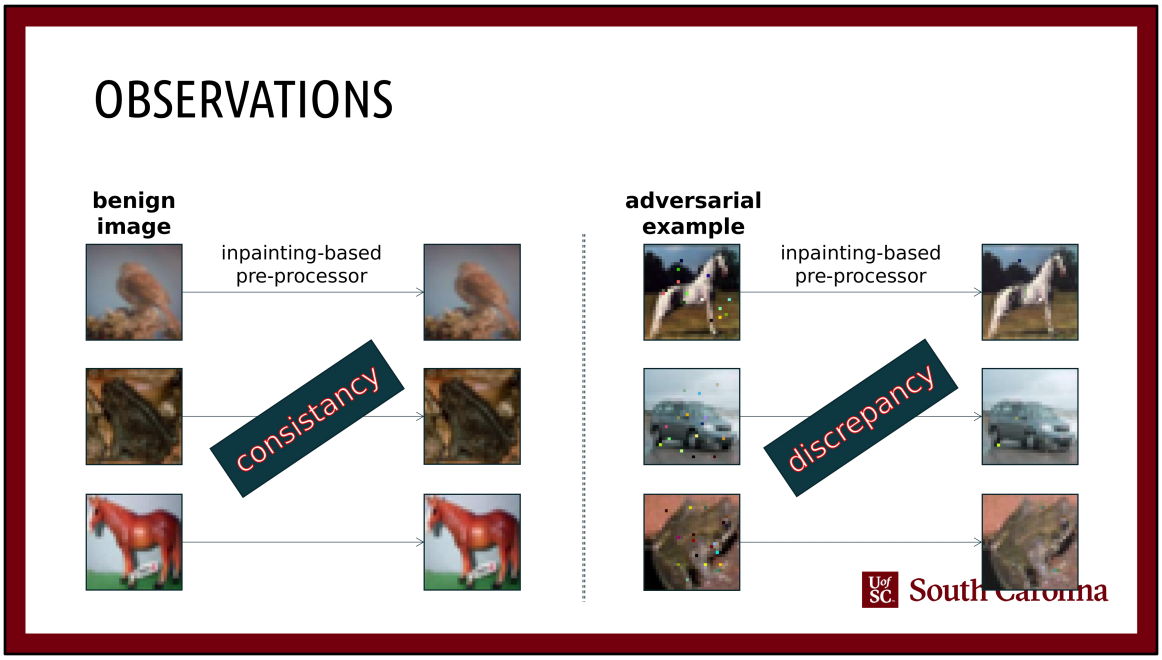
South Carolina

We also can observe a similar result in MNIST datastet.
Note the algorithm for gray images is very similar to the version for color images, but we only need to consider one channel rather than three.

A SIAMESE-NETWORK-BASED AE DETECTOR

South Carolina

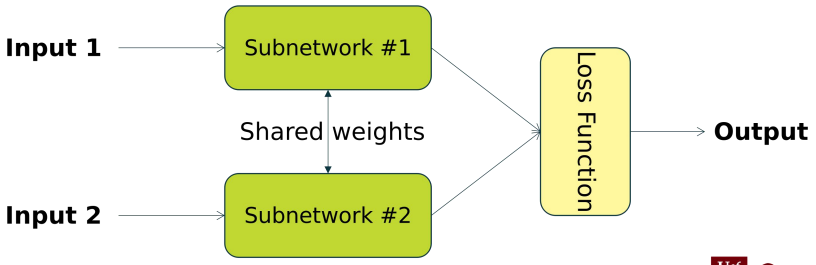
Based on the inpainting-based pre-processor, next we will discuss our detector design.



For a benign image, before and after using our inpainting-based pre-processor, it tends to remain the same.
However, for an L0 AE, before and after using our inpainting-based pre-processor, the image changes to some degree.
We expect an automatic approach to capture the consistancies and the discrepancies.
Fortunately, a Siamese network is capable of this task.

SIAMESE NEURAL NETWORK

- *Siamese neural network* is a class of neural network architectures that consist of two identical subnetworks.

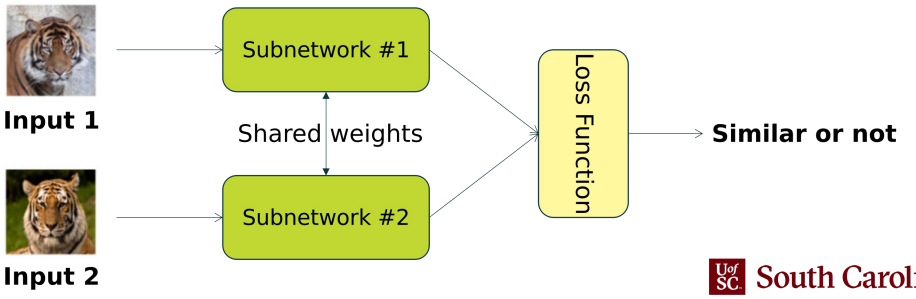


South Carolina

Identical here means they have the same configuration with the same parameters and weights. Parameter updating is mirrored across both subnetworks.

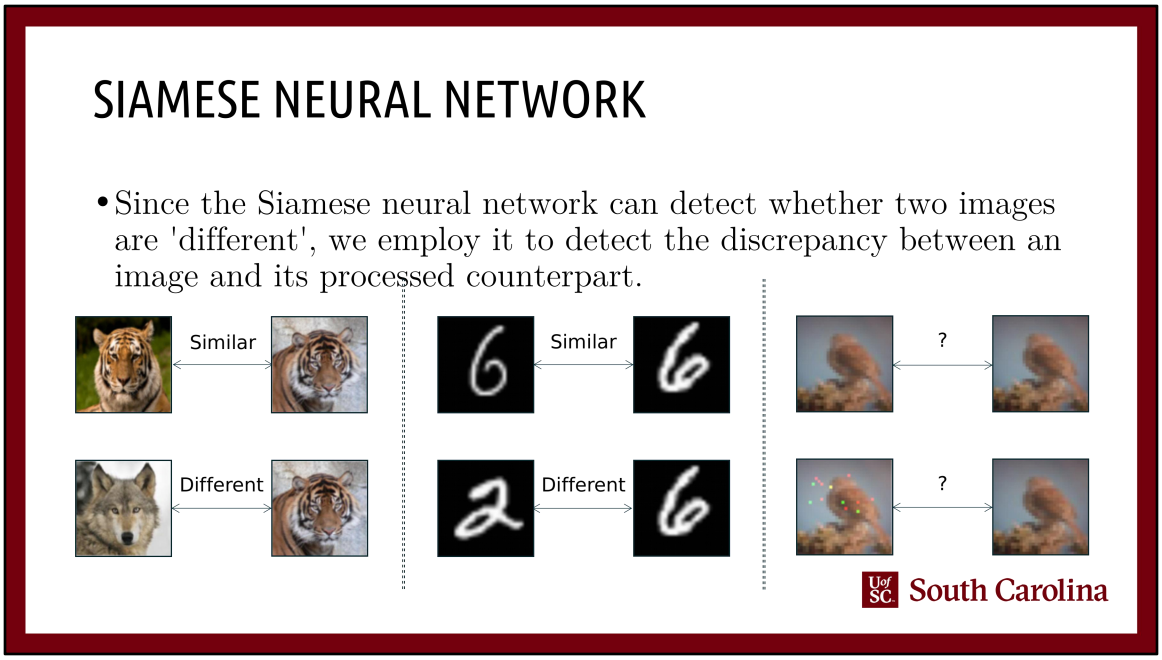
SIAMESE NEURAL NETWORK

- Siamese neural networks are very popular among tasks that involve finding similarity or a relationship between two comparable things.



South Carolina

Take the application in computer vision as an example, each subnetwork takes one of the two input images. The last layers of the two subnetworks are then fed to a contrastive loss function , which calculates the similarity between the two images.



For example, Siamese neural network can successfully assert these two images are both tigers.
It also can correctly state that a wolf is different from a tiger.
Similarly, Siamese neural network can successfully detect whether two hand-written digits are different or not.

If the discrepancy between two images are large enough, we consider the input image as an AE.

SIAMESE-NETWORK-BASED DETECTOR

- Given an image I, we first process it with our inpainting-based algorithm to obtain another image I'.
- The well trained Siamese network takes I and I' as the inputs. It is able to automatically and precisely capture the discrepancies between the two inputs. Lastly, it outputs whether I is an AE.

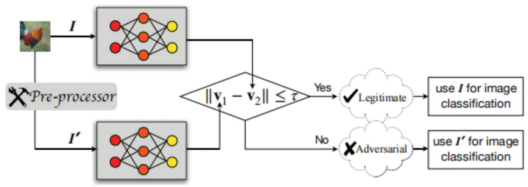
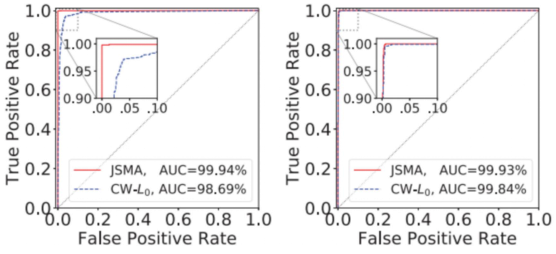


Figure: System architecture

South Carolina

EVALUATION ON THE DETECTION PERFORMANCE

- Our system demonstrates not only high AE detection rate but also low false positive rate.
- For CIFAR-10, it can achieve the AUC values of 98.69% and 99.94% for the two L_0 attacks.
- For MNIST, the AUC value can achieve 99.84% and 99.93%.



(a) CIFAR-10 (b) MNIST

Figure: ROC curves for different datasets

RESILIENCE TO ADAPTIVE ATTACK

South Carolina

Finally, we also consider the scenario of attaptive attacks.
To this end, we assume there exists an adversary who knows the details of our detector
and will try to adapt the attacks accordingly.

ADAPTIVE ATTACK: METHOD

- We launch an adaptive L_0 attack by adopting a similar method described in [He et al., 2017] based on the exploration multiple optimization paths.
- To generate L_0 AEs, after each step of stochastic gradient descent (SGD), an intermediate distorted image is generated as a resolution of the optimizer.

ADAPTIVE ATTACK: METHOD

- Each time the optimizer runs, the process tries to minimizes the number of altered pixels and, in the meanwhile, keep the targeted attack successful.
- We then check whether the intermediate image can bypass our detector. For each image, we repeat the optimization procedure for multiple times to explore different optimization paths (for this purpose, we set a randomly initialized state at the beginning of each optimization procedure).

ADAPTIVE ATTACK: EXPERIMETAL CONCLUSION

- If set the L_0 constraint as the optimization target, it is difficult to control the amplitude of the altered pixels.
- It is challanging for an attacker to adatpively generate L_0 AEs to bypass our detector (the success ratio is only 7%).

TAKEAWAY MESSAGES

- By identifying and exploiting the inherent characteristics of L_0 AEs, we develop a countermeasure that thwarts this type of attacks.
- Its novel Siamese network based design shows very high accuracies in detecting L_0 AEs, and its inpainting-based preprocessing technique can effectively rectify those AEs and thus correct the classification results.
- Last but not least, only controlling the number of altered pixels without limiting the resulting amplitude weakens the power of the generated AEs. Thus, how to make a good trade-off between the number of altered pixels and their amplitude becomes critical when designing new AE generation algorithms.

RESEARCH RESOURCE IS AVAILABLE

- We open-sourced the data and model at the following link:

https://github.com/fzuo/AEPecker

REFERENCES

- Telea, A., 2004. An image inpainting technique based on the fast marching method. Journal of graphics tools, 9(1).
- Elad, M., Starck, J.L., Querre, P. and Donoho, D.L., 2005. Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA). Applied and Computational Harmonic Analysis, 19(3).
- Liang, B. , Li, H., Su, M., Li, X., Shi, W., and Wang, X., 2018. Detecting adversarial image examples in deep neural networks with adaptive noise reduction. IEEE Transactions on Dependable and Secure Computing.
- Xu, W., Evans, D., and Qi, Y., 2018. Feature squeezing: Detecting adversarial examples in deep neural networks. In Network and Distributed System Security Symposium.
- He, W., Wei, J., Chen, X., Carlini, N., and Song, D., 2017. Adversarial example defenses: Ensembles of weak defenses are not strong. In USENIX Workshop on Offensive Technologies.

THANKS!