UVA CS 6316: Machine Learning : 2019 Fall
Course Project: Deep2Reproduce @
https://github.com/qiyanjun/deep2reproduce/tree/master/2019Fall

Decision Boundary Analysis of Adversarial Examples

Reproduced by: Xugui Zhou

Dec 5th, 2019

Motivation

- Deep neural networks (**DNNs**) **are vulnerable** to adversarial examples, which are carefully crafted instances aiming to cause prediction errors for DNNs.

- Recent defensing technique on adversarial examples is **not enough**: examining local neighborhoods in the input space of DNN models, previous work has limited what regions to consider, focusing either on low-dimensional subspaces or small balls.

Background

- **Adversarial examples:** are slightly perturbed versions of correctly classified input instances, which are misclassified.
- The amount of perturbation used to generate an adversarial example from the original input instance is called the example's **distortion**.

Defense against adversarial examples:

- **Adversarial training** with examples generated by projected gradient descent (PGD);
- **Region classification**, takes the majority prediction on several slightly perturbed versions of an input, uniformly sampled from a hypercube around it. In contrast, classifying only the input instance can be referred to as **point classification**.

Related Work

- Liu et al. (2017) and Tramèr et al. (2017) examine limited regions around benign samples to study why some adversarial examples transfer across different models.
- Madry et al. (2017) explore regions around benign samples to validate the robustness of an adversarialy trained model.
- Tabacof& Valle(2016) examine regions around adversarial examples to estimate the examples' robustness to random noise.
- Cao & Gong (2017) determine that considering the region around an input instance produces more robust classification than looking at the input instance alone as a single point.

Limitations:

- focus on low-dimensional subspaces around a model's gradient direction.
- explore many directions, but they focus on a small ball.

Claim / Target Task

Information from larger neighborhoods—both in more directions and at greater distances—will better help us understand adversarial examples in high-dimensional datasets.

An Intuitive Figure Showing WHY Claim

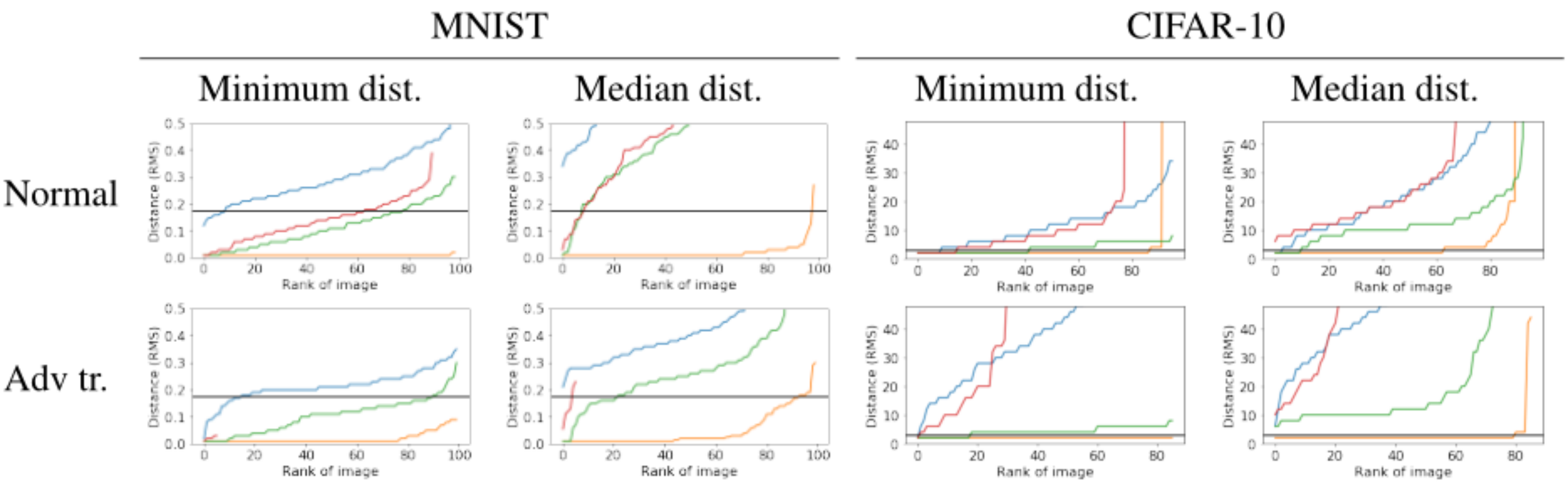


Figure 2: Minimum and median decision boundary distances across random directions, for a sample of images. **Blue**: Benign. **Red**: FGSM. **Green**: OPTMARGIN (ours). **Orange**: OPTBRITTLE. Each statistic is plotted in ascending order. A black line is drawn at the expected distance of images sampled by region classification.

No simple threshold on any one of these statistics accurately separates benign examples (blue) from OPT MARGIN examples (green).

Proposed Solution

- Demonstrate OPT-MARGIN, a new attack that evades region classification systems with low-distortion adversarial examples.
- Analyze a larger neighborhood around input instances by looking at properties of surrounding decision boundaries, namely the distances to the boundaries and the adjacent classes.
- Train a classifier to differentiate the decision boundary information that comes from different types of input instances

Implementation

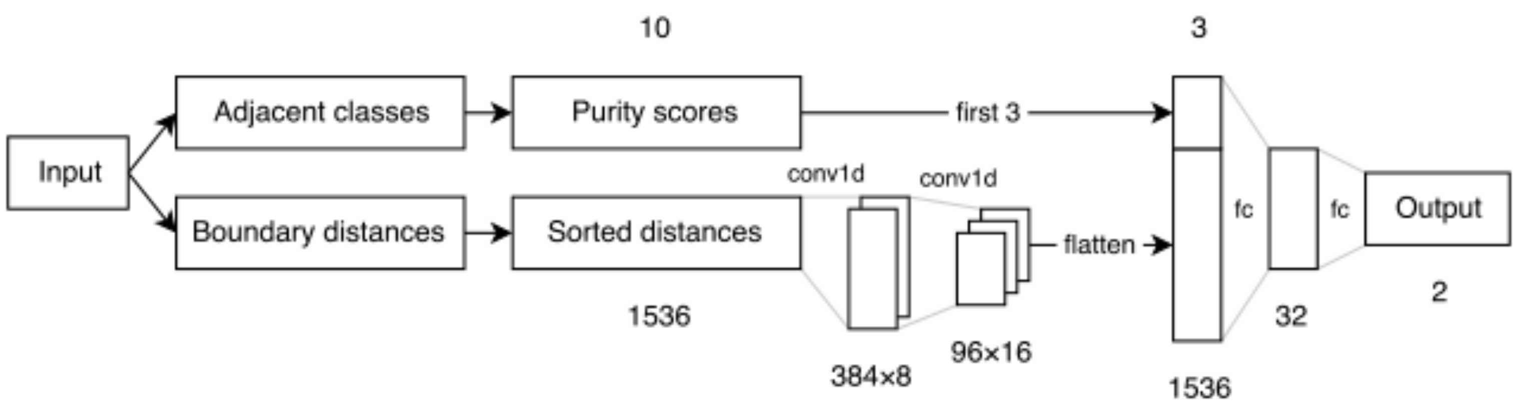
Dataset:

- MNIST, consisting of black-and-white handwritten digits (LeCun, 1998)
- CIFAR-10, consisting of small color pictures (Krizhevsky & Hinton, 2009)
- *a small subset of ImageNet (additionally)*

Model Training:

- MNIST: CNN, both normal and with PGD -L∞ perturbation limit of 0.3
- CIFAR-10: ResNet, bot normal and with PGD -L∞ perturbation limit of 8

New d



Opt-margin Approach

$$\text{minimize} \quad ||x' - x||_2^2 + c \cdot (\ell_1(x') + ... + \ell_n(x'))$$

Let $Z(x)$ refer to the $|C|$-dimensional vector of class weights, in logits, that f internally uses to classify image x. As in Carlini & Wagner's L_2 attack (2017b), we define a loss term for each model in our ensemble:

$$\ell_i(x') = \ell(x' + v_i) = \max\left(-\kappa,\ Z(x' + v_i)_y - \max\{Z(x' + v_i)_j : j \neq y\}\right)$$

v 20 = 0

Data Summary

Examples	MNIST Normal		MNIST Adv tr.		CIFAR-10 Normal		CIFAR-10 Adv tr.	
OPTBRITTLE	100%	0.0732	100%	0.0879	100%	0.824	100%	3.83
OPTMARGIN (ours)	100%	0.158	100%	0.168	100%	1.13	100%	4.08
OPTSTRONG	100%	0.214	28%	0.391	100%	2.86	73%	37.4
FGSM	91%	0.219	6%	0.221	82%	8.00	36%	8.00

Table 1: Success rate (%) and average distortion (RMS) of adversarial examples generated by different attacks. On MNIST, the level of distortion in OPTMARGIN examples is visible to humans, but the original class is still distinctly visible (see Figure 5 in the appendix for sample images).

Examples	MNIST Region cls. Normal	MNIST Region cls. Adv. tr.	MNIST Point cls. Normal	MNIST Point cls. Adv. tr.	CIFAR-10 Region cls. Normal	CIFAR-10 Region cls. Adv. tr.	CIFAR-10 Point cls. Normal	CIFAR-10 Point cls. Adv. tr.
Benign	99%	100%	99%	100%	93%	86%	96%	86%
FGSM	16%	54%	9%	94%	16%	55%	17%	55%
OPTBRITTLE	95%	89%	**0%**	**0%**	71%	79%	**0%**	**0%**
OPTMARGIN (ours)	**1%**	**10%**	**0%**	**0%**	**5%**	**5%**	**0%**	6%

Table 2: Accuracy of region classification and point classification on examples from different attacks. More effective attacks result in lower accuracy. The attacks that achieve the lowest accuracy for each configuration of defenses are shown in bold. We omit comparison with OPTSTRONG due to its disproportionately high distortion and low attack success rate.

Data Summary

Table ... by dif-
feren ... umans,
but th ...).

4

2

0
R

Table 2. Accuracy of region classification and point classification on examples from different at-tacks. More effective attacks result in lower accuracy. The attacks that achieve the lowest accuracy for each configuration of defenses are shown in bold. We omit comparison with OptStrong due to its disproportionately high distortion and low attack success rate.

Experimental Analysis

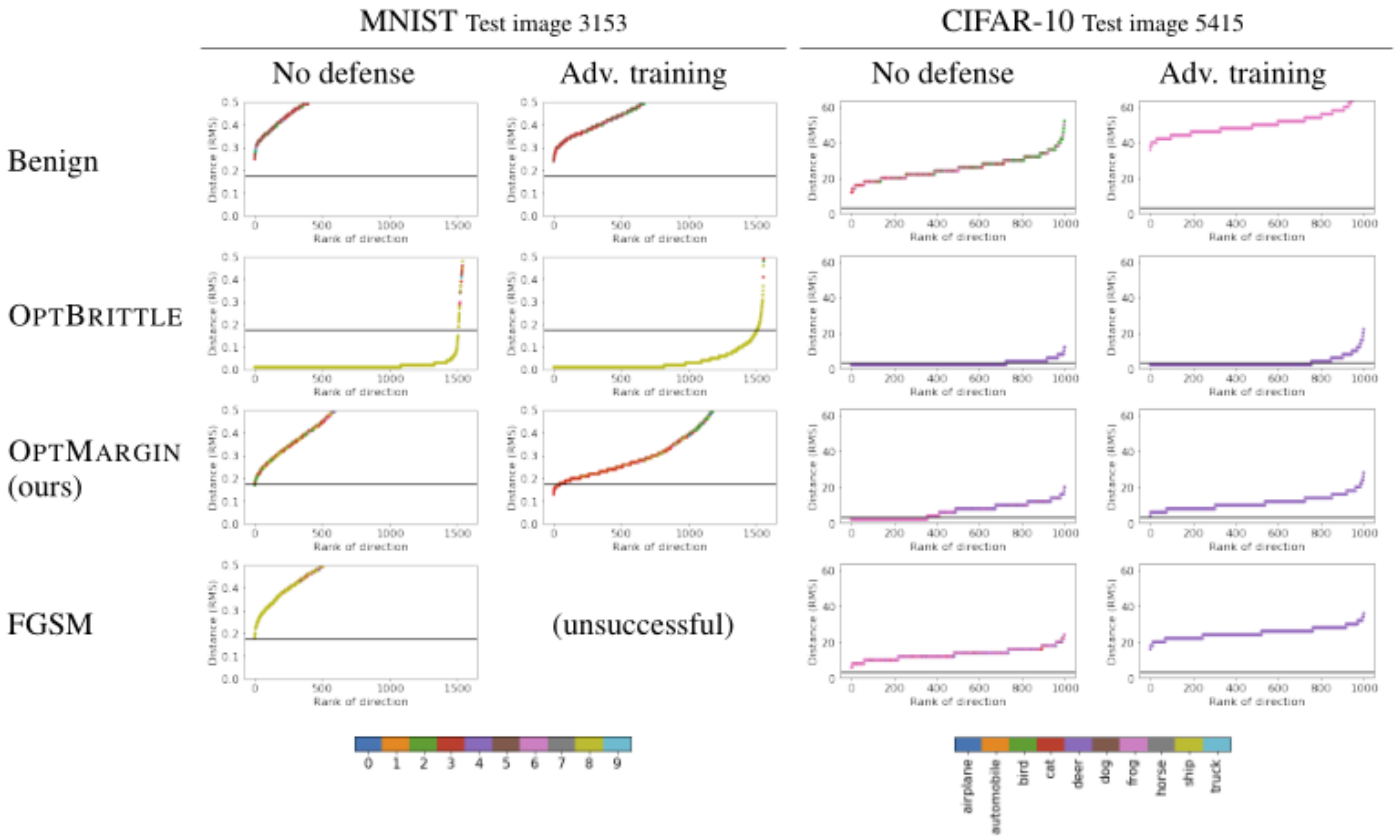


Figure 1: Decision boundary distances (RMS) from single sample images, plotted in ascending order. Colors represent the adjacent class to an encountered boundary. A black line is drawn at the expected distance of an image sampled during region classification. Results are shown for models with normal training and models with PGD adversarial training. For MNIST, original example correctly classified 8 (yellow); OPTBRITTLE and OPTMARGIN examples misclassified as 5 (brown); FGSM example misclassified as 2 (green). For CIFAR-10, original example correctly classified as DEER (purple); OPTBRITTLE, OPTMARGIN, and FGSM examples misclassified as HORSE (gray).

Experimental Analysis

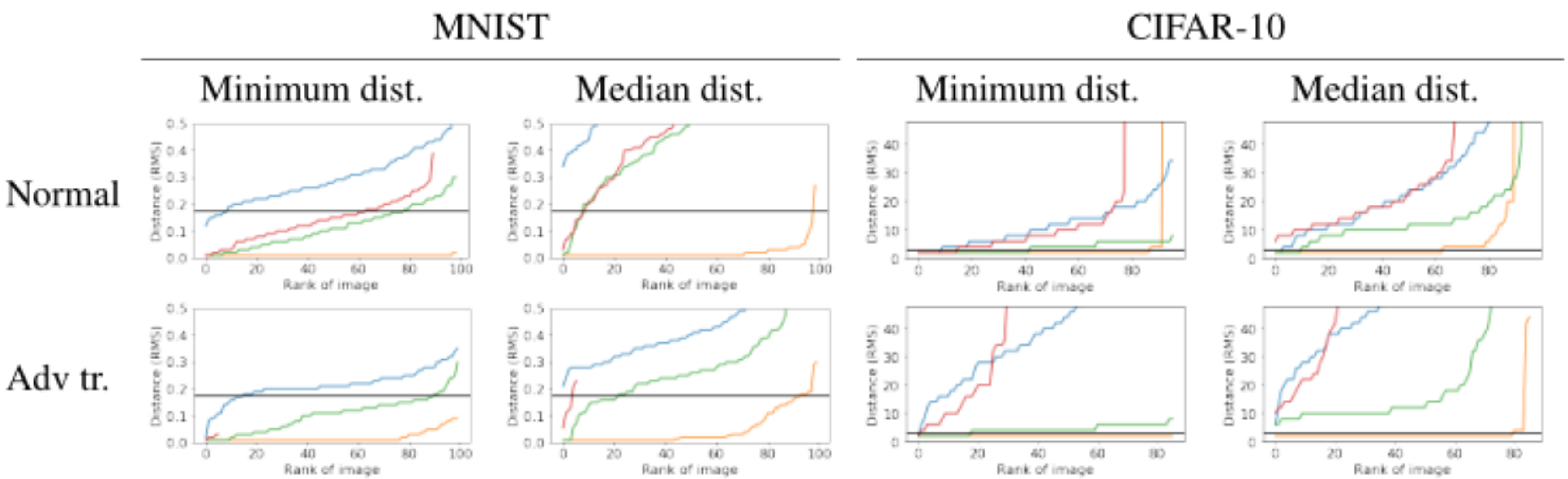


Figure 2: Minimum and median decision boundary distances across random directions, for a sample of images. **Blue**: Benign. **Red**: FGSM. **Green**: OPTMARGIN (ours). **Orange**: OPTBRITTLE. Each statistic is plotted in ascending order. A black line is drawn at the expected distance of images sampled by region classification.

- No simple threshold on any one of these statistics accurately separates benign examples (blue) from OPTMARGIN
- The effect of PGD adversarial training on the robustness of benign examples to random perturbations is not universally beneficial nor harmful.

Experimental Analysis

Adversarial examples generated by OPT MARGIN and FGSM are much harder to distinguish from benign examples in this metric.

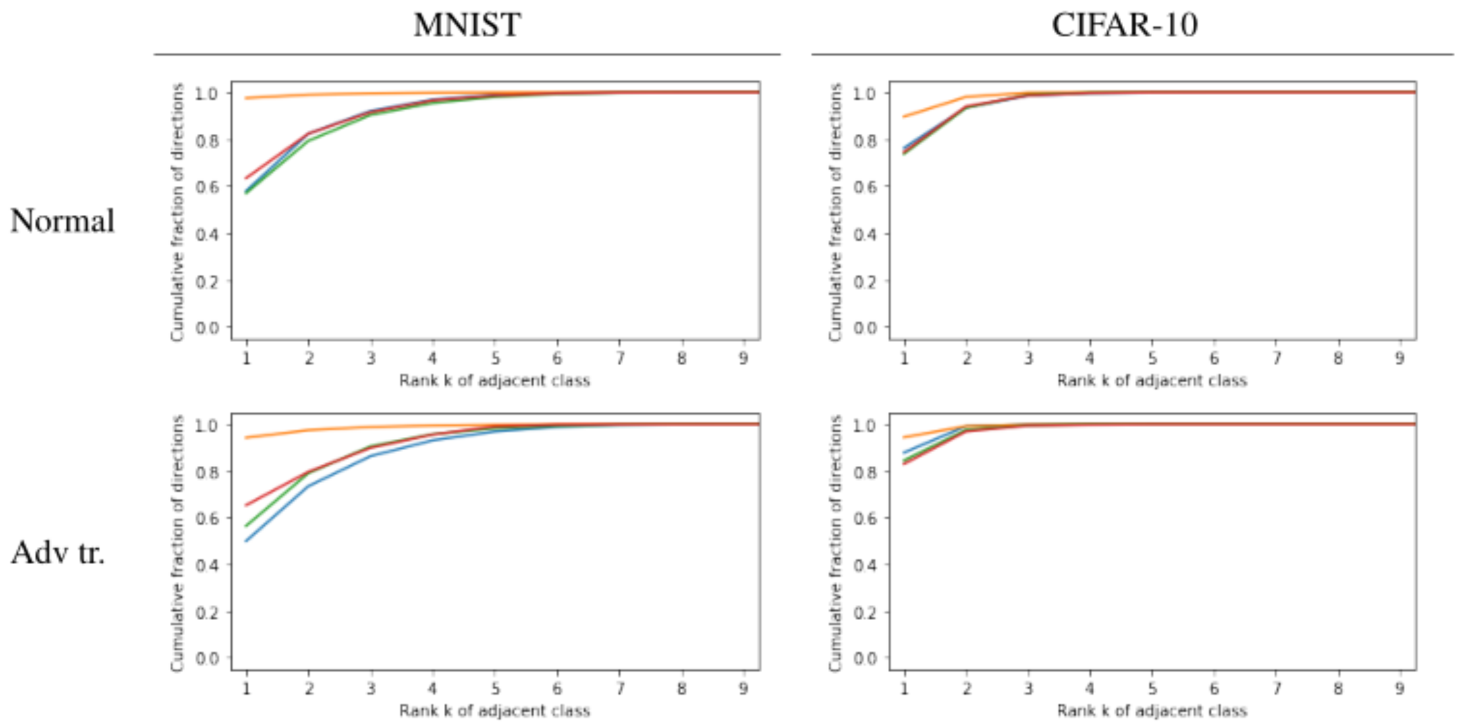


Figure 3: Average purity of adjacent classes around benign and adversarial examples. **Orange**: OPTBRITTLE. **Red**: FGSM. **Green**: OPTMARGIN (ours). **Blue**: Benign. Curves that are lower on the left indicate images surrounded by decision regions of multiple classes. Curves that near the top at rank 1 indicate images surrounded almost entirely by a single class.

Experimental Results

Training attack	False pos. Benign	False neg. OPTBRITTLE	False neg. OPTMARGIN	Accuracy Our approach	Accuracy Cao & Gong
	MNIST, normal training				
OPTBRITTLE	1.0%	1.0%	74.1%	90.4%	10%
OPTMARGIN	**9.6%**	0.6%	7.2%		
	MNIST, PGD adversarial training				
OPTBRITTLE	2.6%	2.0%	39.8%		
OPTMARGIN	10.3%	0.4%	14.5%		
	CIFAR-10, normal training				
OPTBRITTLE	5.3%	3.2%	56.8%	96.4%	5%
OPTMARGIN	8.4%	7.4%	5.3%		
	CIFAR-10, PGD adversarial training				
OPTBRITTLE	0.0%	2.4%	51.8%		
OPTMARGIN	**3.6%**	0.0%	1.2%		

Table 3: False positive and false negative rates for the decision boundary classifier, trained on examples from one attack and evaluated examples generated by the same or a different attack. We consider the accuracy under the worst-case benign/adversarial data split (all-benign if false positive rate is higher; all-adversarial if false negative rate is higher), and we select the best choice of base model and training set. These best-of-worst-case numbers are shown in bold and compared with Cao & Gong's approach from Table 2.

Experimental Represent

Table 1: Success rate(%) and average distortion of adversarial examples generated by OptMargin attack

	MNIST				CIFAR-10			
	Normal		Adv tr.		Normal		Adv tr.	
OptMargin	100%	0.164	100%	0.165	100%	1.248	100%	4.310

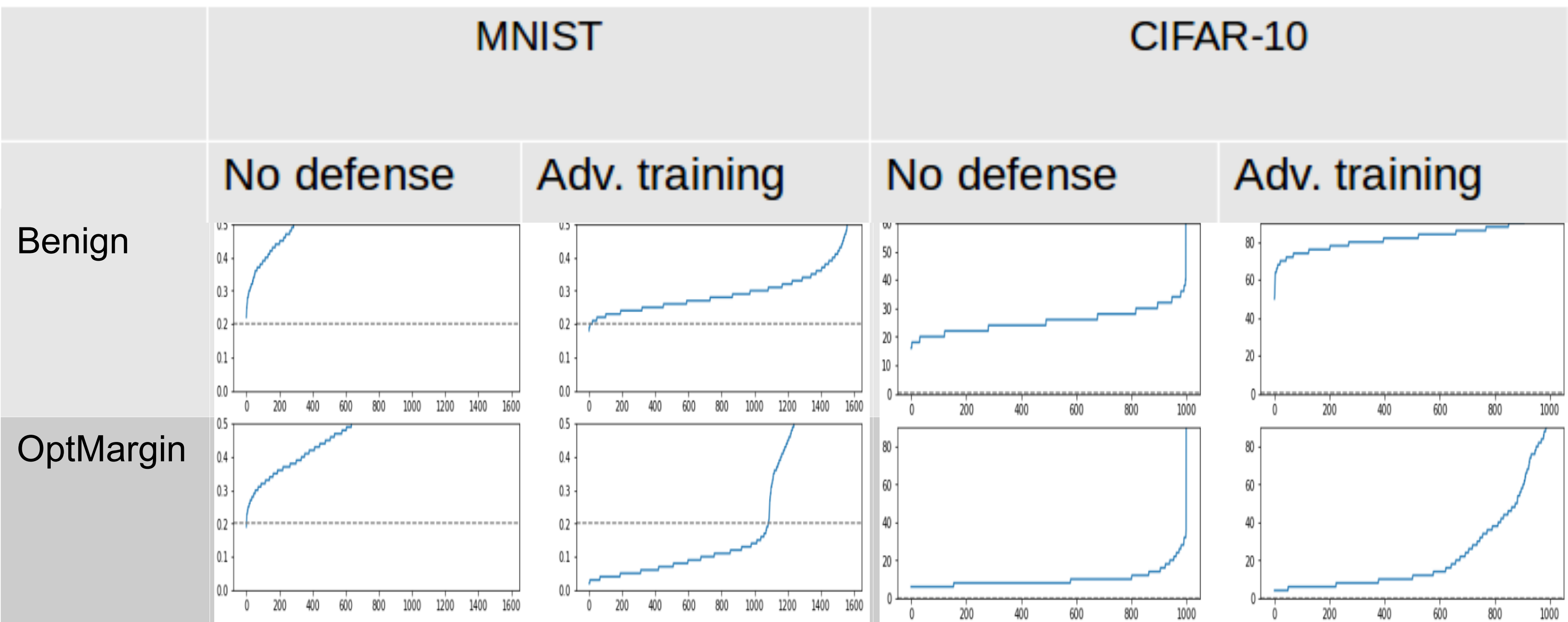
Experimental Represent

Table 2: Accuracy of region classification and point classification

	MNIST				CIFAR-10			
	Region cls.		Point cls.		Region cls.		Point cls.	
	Normal	Adv. tr.	Normal	Adv. tr.	Normal	Adv. tr.	Normal	Adv. tr.
Benign	99%	98%	99%	98%	100%	100%	100%	100%
OptMargin	4%	7%	0%	0%	4.28%	4.78%	4.16%	4.72%

Experimental Represent

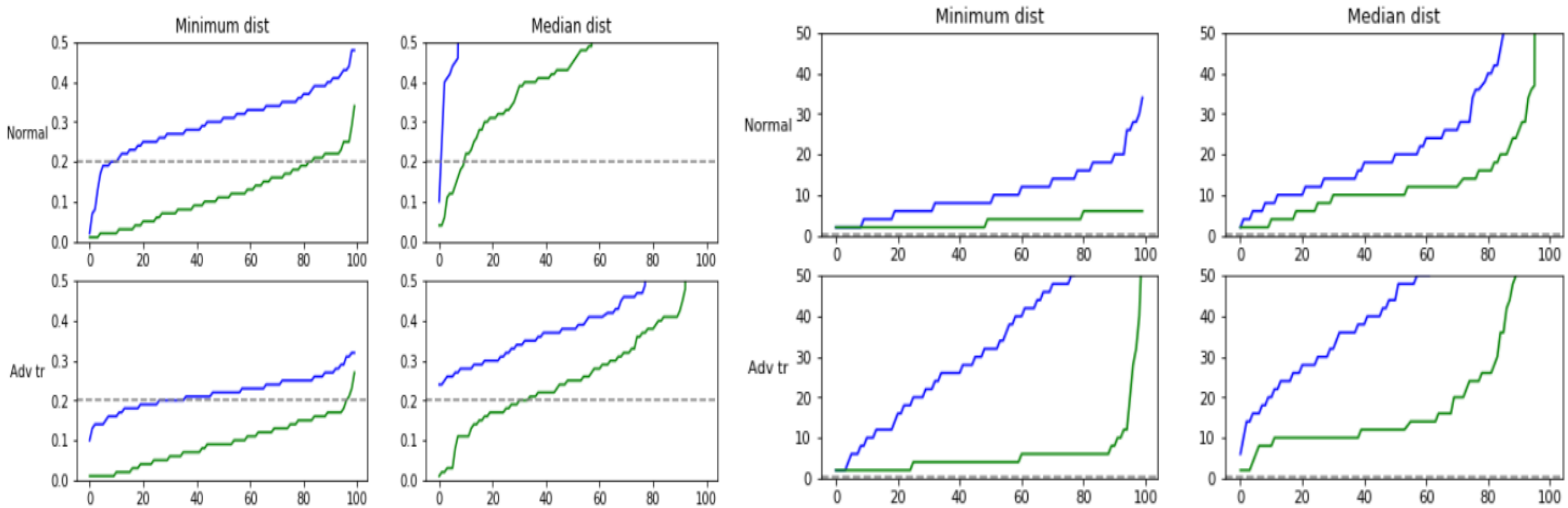
Figure 1: Decision boundary distance from single sample images



Experimental Represent

Figure 2: Minimum and median decision boundary distances for a sample of images: blue(benign), Green(OptMargin)

MNIST	CIFAR10



Conclusion and Future Work

- benefits of examining large neighborhoods around a given input in input space
- We demonstrated an effective OPTMARGIN attack against a region classification defense, which only considered a small ball of the input space around a given instance.
- The comprehensive information about surrounding decision boundaries reveals there are still differences between our robust adversarial examples and benign examples.
- It remains to be seen how attackers might generate adversarial examples that better mimic benign examples' surrounding decision boundaries.

What does each member do?

- Read the paper
- Download the code from github
- Read the code to match the code with the paper
- Represent the experiment (running for more than a month):
  - OptMargin attack
  - Decision boundary analysis
  - *Train a classifier to defend the attack (not achieved)*
- Write scripts to analysis collected experiment data
- Prepare the presentation and jupyter notebook

References

- Xiaoyu Cao and Neil Zhenqiang Gong. Mitigating evasion attacks to deep neural networks via region-based classification. Annual Computer Security Applications Conference (ACSAC), 2017.
- Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. ACM Workshop on Artificial Intelligence and Security (AISEC), 2017a.
- Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In Security and Privacy (SP), 2017 IEEE Symposium on, pp. 39–57. IEEE, 2017b.
- Reuben Feinman, Ryan R Curtin, Saurabh Shintre, and Andrew B Gardner. Detecting adversarial samples from artifacts. arXiv preprint arXiv:1703.00410, 2017.
- Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. 3rd International Conference on Learning Representations (ICLR), 2015.
- Warren He, James Wei, Xinyun Chen, Nicholas Carlini, and Dawn Song. Adversarial example defense: Ensembles of weak defenses are not strong. In 11th USENIX Workshop on Offensive Technologies (WOOT 17), Vancouver, BC, 2017. USENIX Association. URL https://www.usenix.org/conference/woot17/workshop-program/presentation/he.
- Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdi-nov. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580, 2012.

References

- Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.
- Yann LeCun. The MNIST database of handwritten digits. 1998. URL http://yann.lecun.com/exdb/mnist/.
- Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial exam-ples and black-box attacks. 5th International Conference on Learning Representations (ICLR),2017.
- Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu.Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083,2017.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV), 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.
- Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In AAAI, pp. 4278–4284, 2017.
- Pedro Tabacof and Eduardo Valle. Exploring the space of adversarial images. In Neural Networks (IJCNN), 2016 International Joint Conference on, pp. 426–433. IEEE, 2016.
- Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. The space of transferable adversarial examples. arXiv preprint arXiv:1704.03453, 2017.