Towards Dependable Deep CNNs with Out-distribution Learning

Presented by:
Arezoo Rajabi

Mahdieh Abbasi, Arezoo Rajabi, Christian Gagné, Rakesh B. Bobba

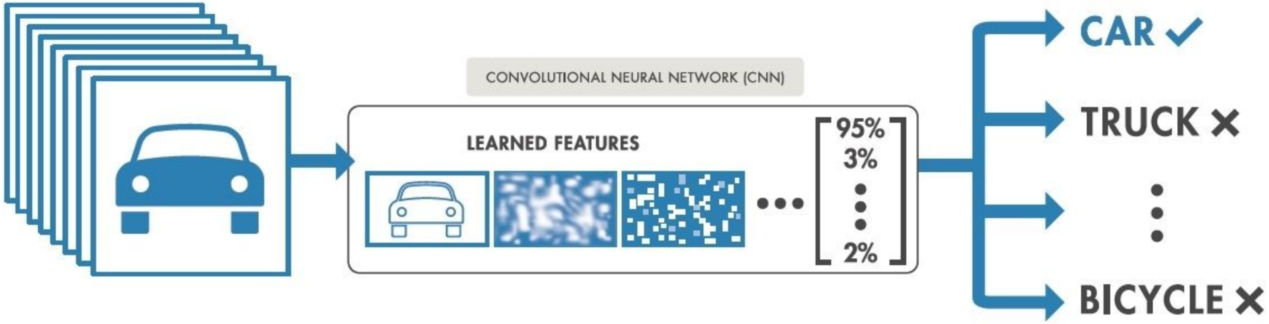



UNIVERSITÉ LAVAL
Oregon State University

Introduction

- Convolutional Neural Networks (CNNs) perform remarkably accurate on large-scale and complex image datasets such as ImageNet

- But they are vulnerable to **adversarial examples** and **out-distribution**

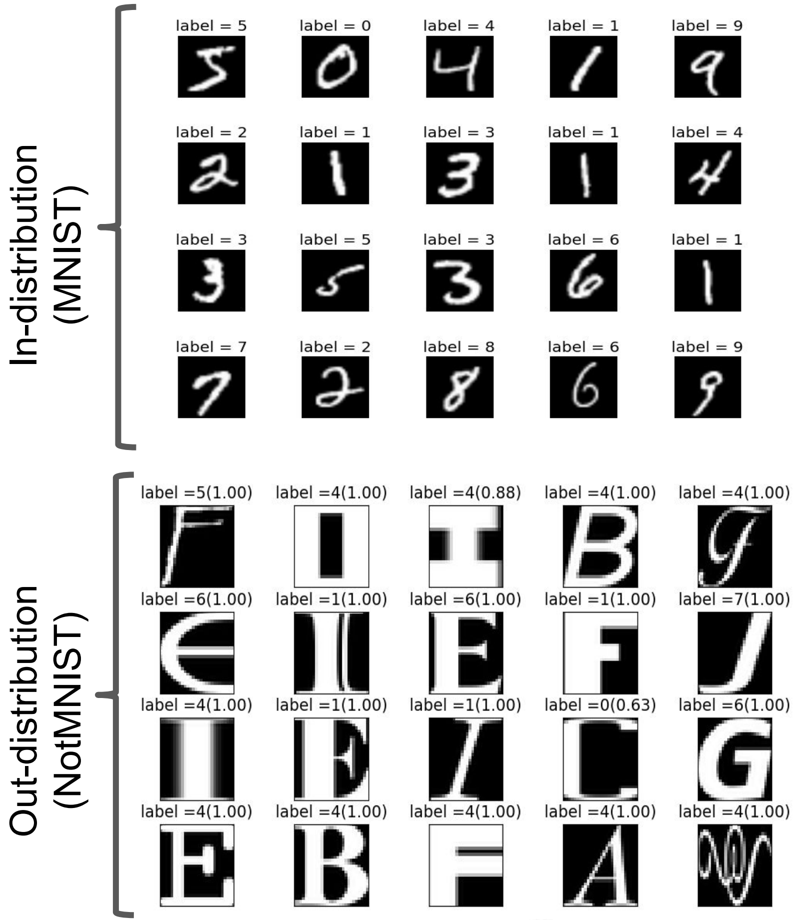


Université Laval

Oregon State University

Out-distribution Problem

- **In-distribution samples**: Samples that belong to the same distribution as training samples
  - CNNs achieve high accuracy on in-dist. samples
- **Out-distribution samples**: Samples not from the same distribution (concept) as training samples
  - CNNs **confidently misclassify** them as one of the trained concepts (classes)



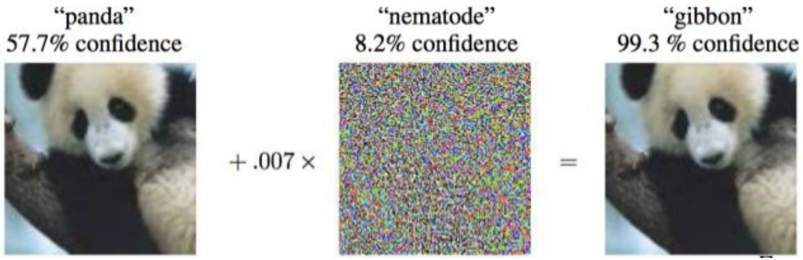
UNIVERSITÉ LAVAL

Oregon State University

Adversarial Examples

- A benign in-dist. sample (x) with wisely added noise (δ) to fool a CNN (F)

$$\min_{\delta} ||\delta||_p$$
$$s.t. \quad F(x+\delta) \neq y^*$$



Adversarial Example [Goodfellow2014ICLR]

- Black-box attack:
  - Learning adversarial samples on a local CNN to attack other victim CNNs

- White-box attack:
  - Assuming have access to a victim CNN, then attacking it by generating adversaries using the victim CNN

Related Work

- Detection and Rejection:
  - Learning on benign and **adversarial examples** to detect and reject them.
  - Some of them needs to learn additional networks
    - Feature squeezing [Xu2018NDSS] : uses adversarial examples to tune threshold for adversarial detection
    - Gross, et. al. train on various adversarial examples to be classified as dustbin

Related Work (cont.)

- Robust CNNs:
  - Aiming to classify adversarial examples **correctly**
    - Madry, et. al. Learn CNNs over a large number of adversarial examples within an ε-neighboring ball of each benign sample [Madry2018ICLR]
    - Distilled network [Papernot2016S&P] obfuscates the gradient of CNNs to make CNNs robust to white-box attacks. But it has been broken by [Carlini2017S&P]

Contributions

1. Draw a connection between overgeneralization and lack of robustness of CNNs
2. Learning an augmented CNN to simultaneously:
   - detect out-distribution samples
   - reduce misclassification rate of <u>black-box</u> adversarial examples

without

- learning on adversarial examples. Most previous defenses are highly dependent on accessing to a diverse set of adversarial examples
- sacrificing CNN's accuracy significantly
- additional computational overhead

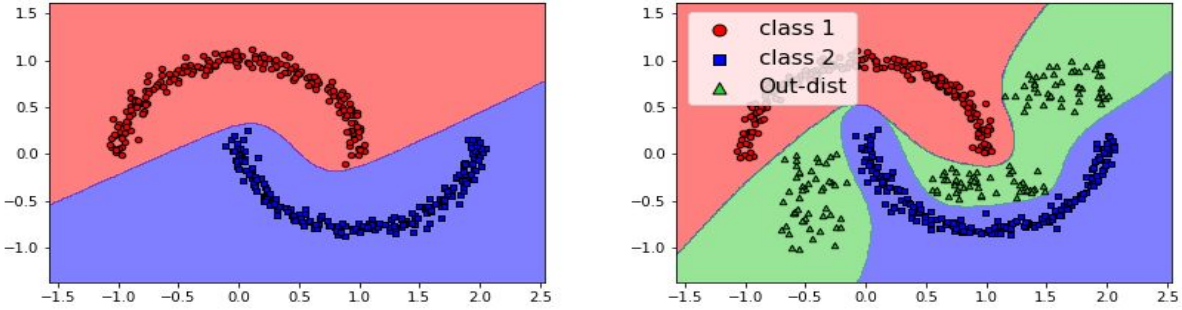
UNIVERSITÉ LAVAL

Oregon State University

Motivation

Reducing overgeneralization of CNNs in out-distribution regions to decrease misclassification rates of adversarial examples and out-dist. samples

Adversarial examples are indeed out-dist. samples[Gross2017arxiv]



Two-moon dataset: (left) decision regions by a naive MLP, (right) decision regions by a augmented MLP.

UNIVERSITÉ LAVAL

Oregon State University

Proposed Approach

We train the augmented CNN on two additional sets of data (along with original in-dist. samples):

1. Out-distribution set:
    - Natural samples available from other task-irrelevant dataset; not (semantically) belonging to in-dist classes

2. Interpolated set:
    - Interpolated samples from pairs of in-dist. samples from two different classes
    - Intuition: an adversarial example contains two different kinds of features
        - visible features related to a true class
        - invisible features related to a fooling class

UNIVERSITÉ LAVAL

Oregon State University

Proposed Approach (cont)

Interpolated samples: For each sample, we selected the nearest samples from other other classes (the images may be misclassified to the source image).

$$I_{int} = \alpha I_{c1} + (1 - \alpha) I_{c2}$$

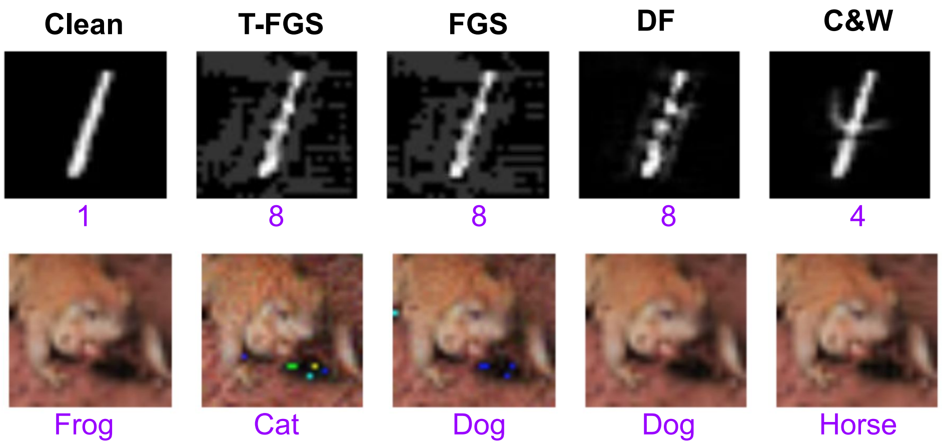
UNIVERSITÉ LAVAL

Oregon State University

Evaluation

Attack Algorithms:

1. Fast Gradient Sign (FGS) [Goodfellow2017]
2. Targeted FGS (T-FGS) [Goodfellow2017]
3. Iterative FGS (I-FGS) [Goodfellow2017]
4. DeepFool (DF) [Moosavi2016]
5. Carlini and Wagner (C&W) [Carlini2017]



Four types of adversarial examples for MNIST (first row) and CIFAR-10 (second row)

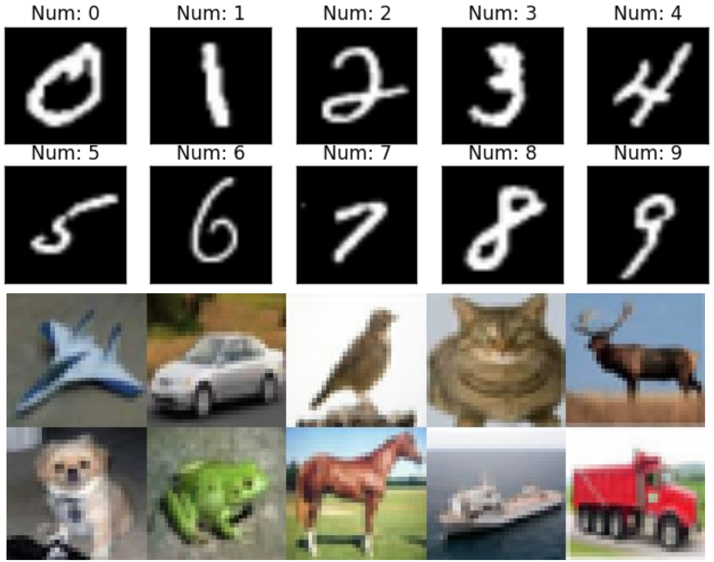
UNIVERSITÉ LAVAL

Oregon State University

Evaluation (cont.)

- Dataset
  - MNIST [LeCun1998]
  - CIFAR-10 [Krizhevsky2009]

- Criteria
  - Accuracy: correct classification rate
  - Rejection: assigned to dustbin class
  - Error: misclassification rate



Evaluation: Black-box MNIST Adversaries

Training set: <in-dist, out-dist.>		<MNIST, —>	<MNIST, NotMNIST>	<MNIST, NotMNIST+intrpl.>
Model		Naive CNN	Augmented CNN	Augmented CNN
In-dist. test	Acc.	**99.5**	99.47	99.48
Out-dist. test	Rej.	-	99.96	**99.98**
FGS	Acc	35.14	19.15	
	Rej	–	65.19	99.59
	Err	64.86	15.66	**0.06**
I-FGS	Acc	16.37	30.97	0.0
	Rej	-	27.08	100
	Err	83.63	41.95	**0.0**
T-FGS	Acc	19.99	1.17	0.0
	Rej	-	95.92	100
	Err	80.01	2.91	**0.0**
DeepFool	Acc	1.89	11.45	5.37
	Rej	-	4.72	89.84
	Err	98.11	83.83	**4.8**
C&W (L_2)	Acc	22.49	27.5	7.5
	Rej	-	5.99	77.49
	Err	77.51	66.51	**15.01**
Average Error		80.82	42.17	**3.97**

Evaluation: Black-box CIFAR-10 Adversaries

Training set: <in-dist, out-dist.>		<CIFAR-10, —>	<CIFAR-10, CIFAR100>	<CIFAR-10, CIFAR100+intrpl.>
Model		Naive VGG	Augmented VGG	Augmented VGG
In-dist. test	Acc.	**90.53**	88.58	86.65
Out-dist. test	Rej.	-	95.36	**96.21**
FGS	Acc	36.16	27.65	23.94
	Rej	-	38.94	49.23
	Err	63.84	33.41	**26.83**
I-FGS	Acc	50.34	45.98	41.92
	Rej	-	18.57	25.88
	Err	49.66	35.45	**32.2**
T-FGS	Acc	36.24	27.06	24.2
	Rej	-	40.54	50.77
	Err	63.76	32.4	**25.03**
DeepFool	Acc	56.82	45.63	42.31
	Rej	-	31.0	38.86
	Err	43.18	23.37	**18.83**
C&W (L_2)	Acc	42.5	46.5	39
	Rej	-	18.5	39.5
	Err	57.5	35	**21.5**
Average Error rate		55.59	31.92	**24.88**

More Expressive Feature Space of Augmented CNN

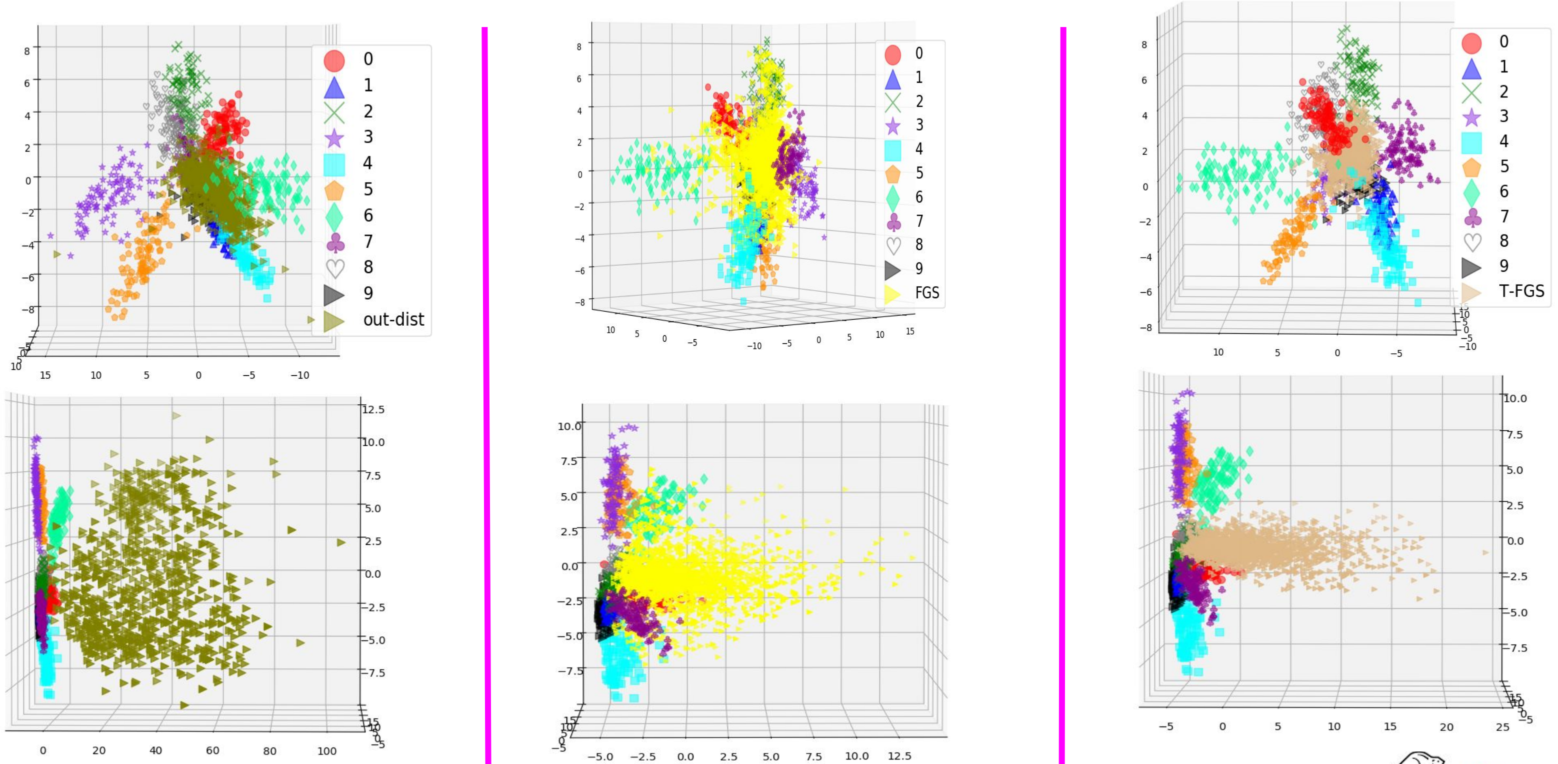
- The penultimate layer of a CNN can be regarded as feature space [benjo2009]
- Augmented CNNs learn more expressive and representative feature spaces such that:
  - **Disentangle natural out-dist. samples from in-dist. ones**
  - **Also separate many of adversaries (without even trained the CNN on them)**

Comparison of feature spaces* - MNIST

Out-dis. samples | FGS adversaries | T-FGS adversaries

Naive CNN

Augmented CNN



UNIVERSITÉ LAVAL

* for visualization purposes, the dimensionality of feature space is reduced to 3 by PCA

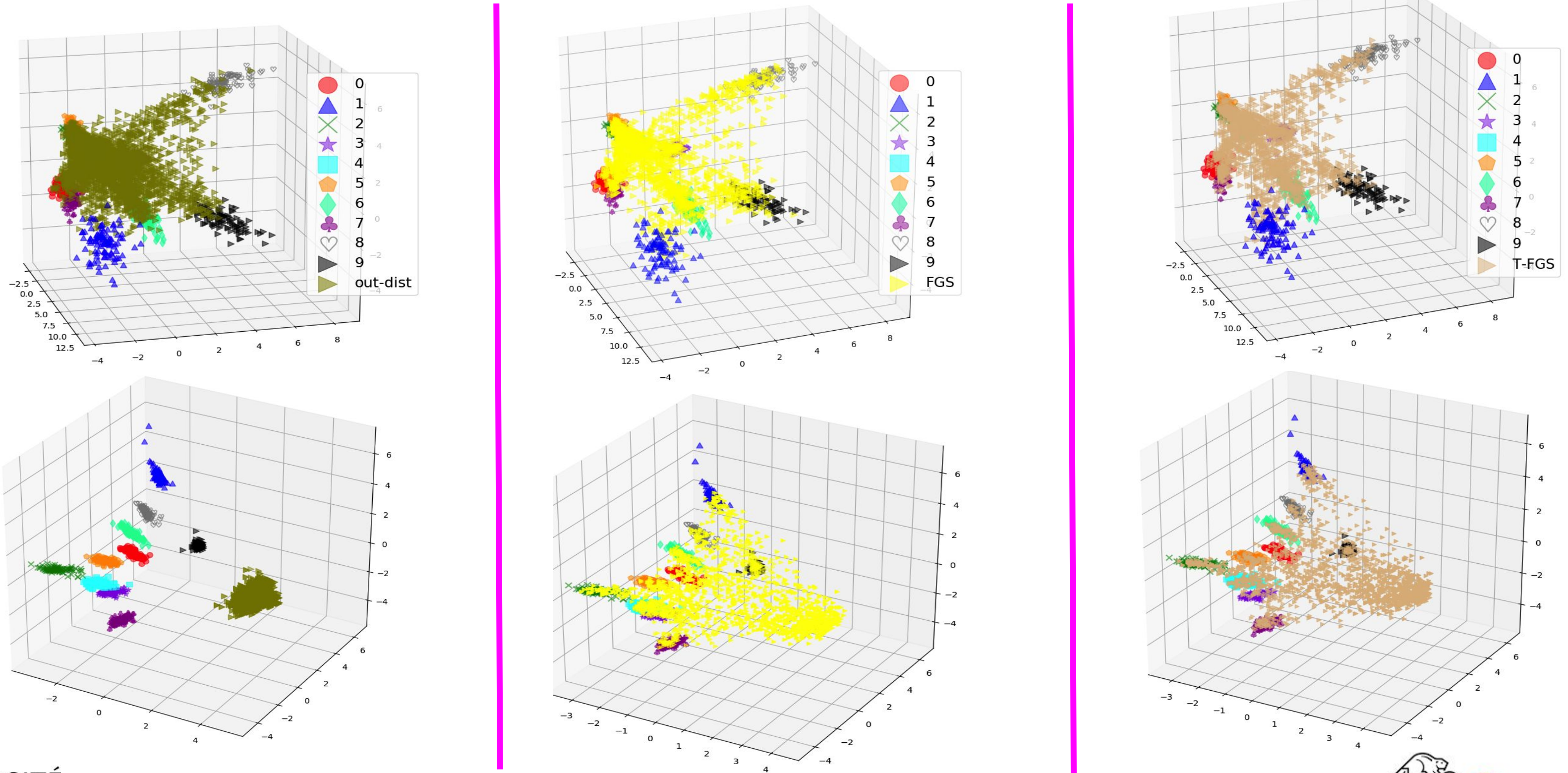
Oregon State University

Comparison of feature spaces - CIFAR10

Out-dis. samples | FGS adversaries | T-FGS adversaries

Naive CNN

Augmented CNN



UNIVERSITÉ LAVAL

Oregon State University

Conclusion

- Augmented CNNs are more **dependable** as they:
  - **Controlling over-generalization** in some out-distribution regions
    - proper decision-making in presence of out-dist. samples by rejecting them as “dustbin”
  - **Distengle some of adversarial examples** from clean samples through learning more expressive feature space
  - **Decreasing error rates** on various types of well-known adversarial examples by **rejecting** them

Future work

- Evaluating augmented CNN in **white-box setting**

- Investigating the features of a an **appropriate** out-dist. sample set

- Evaluating our method on other large-scaled image and non-image datasets

UNIVERSITÉ LAVAL

Oregon State University

Q & A

mahdieh.abbasi.1@ulaval.ca

rajabia@oregonstate.edu

Christian.Gagne@gel.ulaval.ca

rakesh.bobba@oregonstate.edu

UNIVERSITÉ LAVAL

Oregon State University

Reference

[Grosse et al. Arxiv2017] Grosse, K., Manoharan, P., Papernot, N., Backes, M., & McDaniel, P. (2017). On the (statistical) detection of adversarial examples. *arXiv preprint arXiv:1702.06280*.

[Xu et al. NDSS 2018] Xu, W., Evans, D., & Qi, Y. (2018). Feature squeezing: Detecting adversarial examples in deep neural networks. *Network and Distributed System Security Symposium.*

[Papernot et al. S&P2016] Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016, May). Distillation as a defense to adversarial perturbations against deep neural networks. In *Security and Privacy (SP), 2016 IEEE Symposium on* (pp. 582-597).

[Madry et al. ICLR2018] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. International Conference on Learning Representations(ICLR).

[Carlini et al. S&P2017] Carlini, N., & Wagner, D. (2017, May). Towards evaluating the robustness of neural networks. In *Security and Privacy (SP), 2017 IEEE Symposium on* (pp. 39-57). IEEE.

Reference

[Kurakin et al. ICLR2017] Kurakin, A., Goodfellow, I., & Bengio, S. (2017). Adversarial examples in the physical world. *International Conference on Learning Representations*, 2017

[Moosavi et al. CVPR2016] Moosavi-Dezfooli, S. M., Fawzi, A., & Frossard, P. (2016). Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2574-2582).

[LeCun 1998] Y. LeCun. The mnist database of handwritten digits. *http://yann.lecun. com/exdb/mnist/*, 1998.

[Krizhevsky 2009] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. 2009.

[Bengio 2009] Bengio, Yoshua. "Learning Deep Architectures for AI." Machine Learning 2.1 (2009): 1-127.

[Goodfellow et al. ICLR2014] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). International Conference on Learning Representations (ICLR).