Adversarial examples
（对抗样本）

2019/11/2

Made by Shen Haojing & Chen Sihong

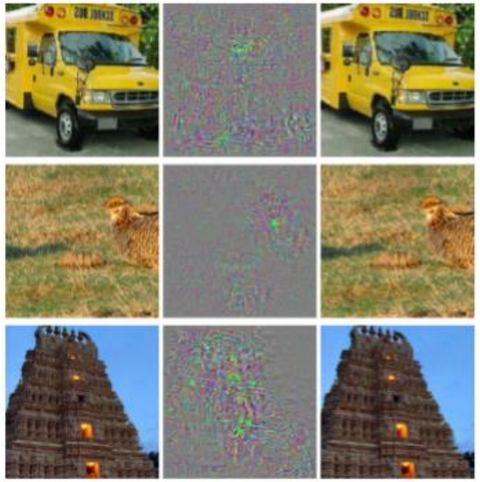
Catalogue

1. What’s adversarial examples?
2. The meaning for studying adversarial examples.
3. Taxonomy of attacks
4. Taxonomy of defenses
5. Challenges in the future

1. What's adversarial examples?

- Adversarial examples (对抗样本) are imperceptible (不可察觉) to human but can easily fool deep neural networks in the testing stage.
- As a box-constrained optimization problem :

$$
\begin{aligned}
\min_{x'} \quad & \|x' - x\| \\
s.t. \quad & f(x') = l', \\
& f(x) = l, \\
& l \neq l', \\
& x' \in [0, 1],
\end{aligned}
$$



(a)

(b)

Szegedy et al. (2014) [19]

Keep imperceptible

Keep fool model

2. The meaning for studying adversarial examples

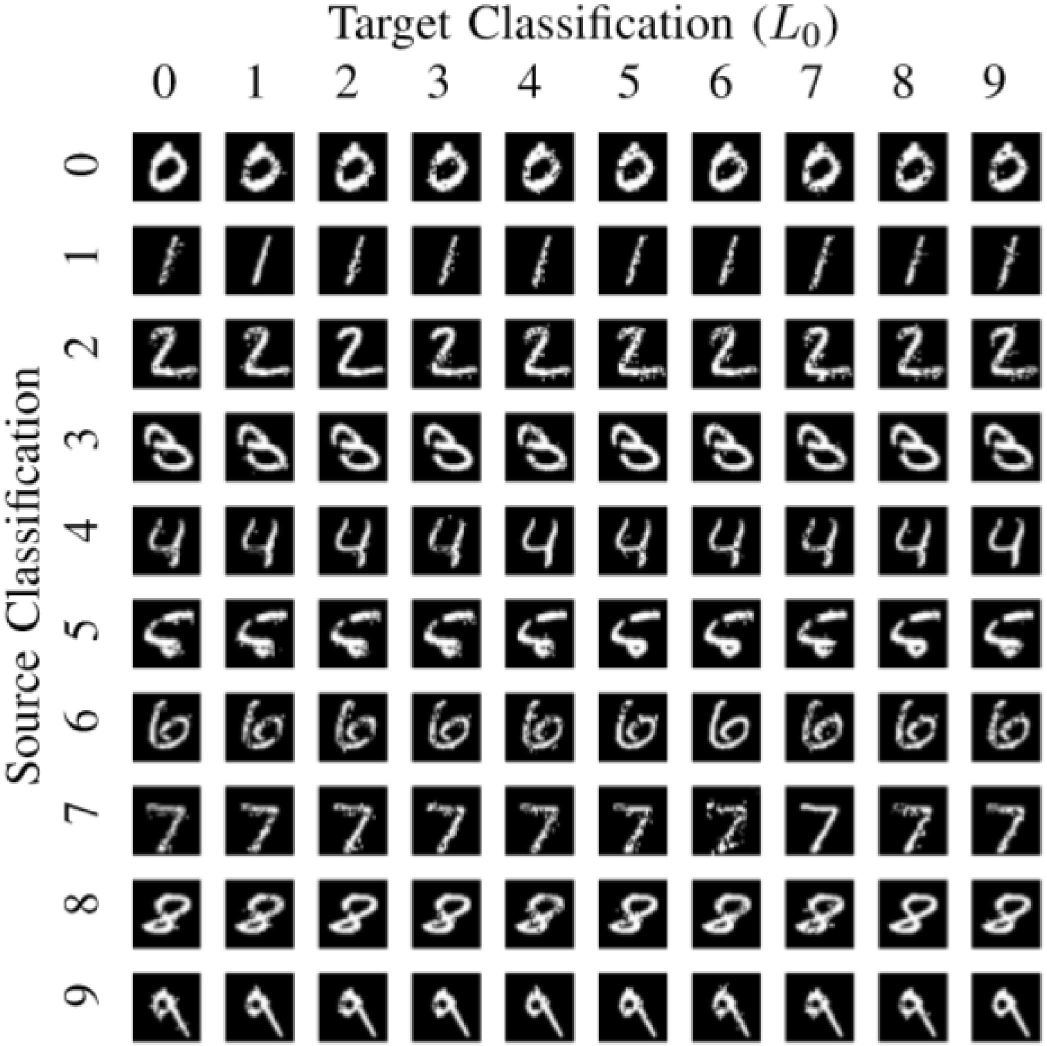
- One of the major risks for applying deep neural networks in safety-critical environments.

- Help us more deeply understand the neural networks. From inspecting adversarial examples, we may gain insights on semantic inner levels of neural networks and problematic decision boundaries.[34]

Help to increase robustness and performance!

3. Taxonomy (分类) of adversarial attacks

- **Adversary's Knowledge**
  1. White-box attacks
  2. Black-box attacks
- **Adversarial Specificity**
  1. Targeted attacks
  2. Non-targeted attacks
- **Attack Frequency**
  1. One-time attacks
  2. Iterative attacks



Adversarial attacks

- **L-BFGS Attack**

  Szegedy et al. firstly introduced adversarial examples against deep neural networks in 2014[19]

- **Fast Gradient Sign Method (FGSM)**

  Goodfellow et al. [69]

- **Basic Iterative Method (BIM) and Iterative Least-Likely Class Method (ILLC) [20]**
- **DeepFool [71]**
- **CPPN EA Fool [83]**
- **C & W's Attack [86]**
- **Zeroth Order Optimization (ZOO) [73]**
- **Universal Perturbation [74]**
- **Feature Adversary [76]**
- **… …**

Adversarial attacks

Applications	Representative Study	Method
Reinforcement Learning	[93]	FGSM
	[94]	FGSM
Generative Modeling	[95]	Feature Adversary, C&W
	[96]	Feature Adversary
Face Recog-nition	[67]	Impersonation & Dodging Attack
Object Detection	[22]	DAG
Semantic Segmentation	[22]	DAG
	[97]	ILLC
	[98]	ILLC
Reading Comprehension	[99]	AddSent, AddAny
	[100]	Reinforcement Learning
Malware Detection	[101]	JSMA
	[102]	Reinforcement Learning
	[103]	GAN
	[104]	GAN
	[105]	Generic Pro-gramming

4. Taxonomy (分类) of Defenses

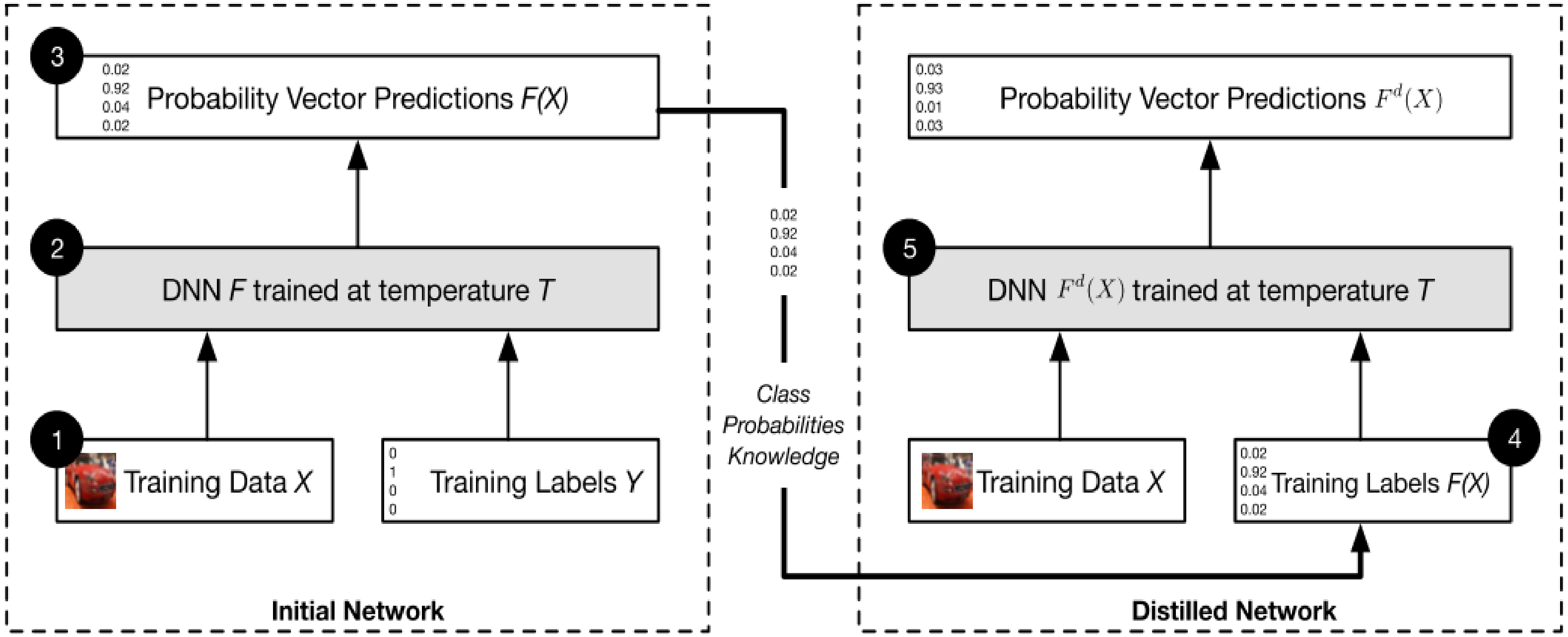
- Network Distillation （蒸馏网络）
- Adversarial training （对抗训练）
- Classifier Robustifying

Defenses

- **Network Distillation （蒸馏网络）**

Network distillation was originally designed to reduce the size of deep neural networks by transferring knowledge from a large networks to a small one [131].

Network distillation extracted knowledge from deep neural networks to improve robustness.[126]



Defenses

- **Adversarial training （对抗训练）**

  Training with adversarial examples is one of the countermeasures to make neural network more robust [69][127].

  Adversarial training increased the robustness of neural networks for one-step attacks (FGSM) but would not help under iterative attacks (BIM and ILLC) [81]

  Adversarial trained models are more robust to white-box adversarial examples than to the transferred examples. [84]

  Ensembling Adversarial Training. [84]

Defenses

- **Classifier Robustifying**

  [128][129] designed robust architectures of deep neural networks to prevent adversarial examples.

5. Challenges in future

1. Transferability （转移性）

- Adversarial examples generated against a neural networks can fool the same neural networks by different dataset. [19]
- Adversarial examples generated against a neural networks can fool other networks with different architectures. [44]

2. The existence of Adversarial examples

- Data incompletion [19, 135, 123, 126]
- Model capability [44, 137, 69, 138, 76, 80]
- No robust model [36, 139, 140]

3. Robustness Evaluation

- Base-line attack
- A methodology for evaluation on the robustness of NN.

References

[19] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfel- low, and R. Fergus, "Intriguing properties of neural networks," arXiv preprint arXiv:1312.6199, 2013.

[34] J. Lu, T. Issaranon, and D. Forsyth, "Safetynet: Detecting and rejecting adversarial examples robustly," ICCV, 2017.

[69] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," arXiv preprint arXiv:1412.6572, 2014.

[20] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," arXiv preprint arXiv:1607.02533, 2016.

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

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

[86] N. Carlini and D. Wagner, "Adversarial examples are not easily detected: Bypassing ten detection methods," AISEC, 2017

[73] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," arXiv preprint arXiv:1708.03999, 2017

References

[74] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, “Univer- sal adversarial perturbations,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[76] S. Sabour, Y. Cao, F. Faghri, and D. J. Fleet, “Adversarial manipulation of deep representations,” Proceedings of the International Conference on Learning Representations (ICLR), 2016.

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

[69] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” arXiv preprint arXiv:1412.6572, 2014.

[127] R. Huang, B. Xu, D. Schuurmans, and C. Szepesvári, “Learning with a strong adversary,” arXiv preprint arXiv:1511.03034, 2015.

[81] A. Kurakin, I. Goodfellow, and S. Bengio, “Adversarial machine learning at scale,” Proceedings of the International Conference on Learning Representations (ICLR), 2017

[84] F. Tramèr, A. Kurakin, N. Papernot, D. Boneh, and P. McDaniel, “Ensemble adversarial training: Attacks and defenses,” arXiv preprint arXiv:1705.07204, 2017.

[128] J. Bradshaw, A. G. d. G. Matthews, and Z. Ghahramani, “Adversarial examples, uncertainty, and transfer testing robustness in gaussian process hybrid deep networks,” arXiv preprint arXiv:1707.02476, 2017.

[129] M. Abbasi and C. Gagné, “Robustness to adversarial examples through an ensemble of specialists,” arXiv preprint arXiv:1702.06856, 2017.