Annotated Reference

Han Xintian

December 5, 2018

1 Adversarial Examples

1.1 Adversarial Attacks

1.1.1 LBFGS Attack

Intriguing properties of neural networks [5]. Given an image x, L-BFGS tries to find a different image x' that is close to x. They solve the following constrained problem to find x':

$$\begin{aligned} \min \quad & \|x-x'\|_2^2 \\ s.t. \quad & C(x') = l \\ & x' \in [0,1]^n, \end{aligned}$$

where l is the target label. The original problem is hard to solve. They solve the following problem instead.

$$\begin{aligned} \min \quad & c \cdot \|x-x'\|_2^2 + \text{loss}_{F,l}(x') \\ s.t. \quad & x' \in [0,1]^n, \end{aligned}$$

where loss function $\text{loss}_{F,l}$ is a function that maps an image to a positive label l, for example, cross-entropy loss. It aims to perform targeted attack.

1.1.2 Fast Gradient Sign Method (FSGM)

Explaining and Harnessing Adversarial Examples [2]. FSGM is a fast algorithm. It will not produce very close adversarial examples. Given an image x, FSGM sets

$$x' = x - \varepsilon \cdot \text{sign}(\nabla \text{loss}_{F,t}(x)),$$

where ε is chosen to be sufficiently small so as to be undetectable, and t is the target label. An improved version is Iterative Gradient Sign. Begin by setting $x'_0 = x$, we have

$$x'_i = Clip_{x,\varepsilon}\{x'_{i-1} + \alpha \text{sign}(\nabla_{F,t}(x'_{i-1}))\}$$

It is an untargeted attack.

1.1.3 JSMA

The limitations of deep learning in adversarial settings [4]. JSMA is short for Jacobian-based Saliency Map Attack. It is still a gradient based attack. They use the gradient $\nabla Z(x)_l$ to compute a saliency map, which models the impact each pixel has on the resulting classification. They choose selected number of pixels changing which will make the target class more likely and other classes less likely.

1.1.4 Deepfool

Deepfool: a simple and accurate method to fool deep neural networks [3]. They image the neural network are generally linear with a hyperplane separating each class from another. They create adversarial examples for this simplified problem and use these examples to attack neural network. They use the adversarial examples to attack another classifier.

1.1.5 Carlini's l_0, l_2 and l_∞ Attack

Defensive distillation is not robust to adversarial examples [1]. The Carlini's attack mainly solve the problem:

$$\begin{aligned} \min \quad & \mathscr{D}(x, x+\delta) \\ s.t. \quad & C(x+\delta) = t \\ & x+\delta \in [0,1]^n \end{aligned}$$

Here, $\mathscr{D}$ could be l_0, l_2 or l_∞. They define the objective function f such that $C(x+\delta) = t$ if and only if $f(x+\delta) \leq 0$. When they find f, they tries to minimize

$$\begin{aligned} \min \quad & \mathscr{D}(x, x+\delta) + c \cdot f(x+\delta) \\ s.t. \quad & x+\delta \in [0,1]^n \end{aligned}$$

LBFGS is a special case of this method.

1.2 Defense Methods

1.2.1 Adversarial Training

Towards Deep Learning Models Resistant to Adversarial Attacks [3]. Defense method over all first order attack? Adversarial training over projected gradient method. Adversarial training over projected gradient method will also be robust to other first order attack method. None first order attack are hard to reach by a first order method even restarting randomly.

Other findings: 1. Large capacity network is more robust. The loss value decreases after adversarial training.

1.3 Other

1.3.1 Sensitivity and generalization in neural networks [?]

Empirically show that smaller generation gap corresponds to lower sensitivity. Use frobenius norm of the Jacobian and number of transitions (curvature of the functions) to characterize sensitivity. Create close to manifold datasets and off manifolds datasets. Close to manifolds datasets by combination of digits from the same class. Off manifolds by random inputs and combination of digits from different class.

References

[1] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. *arXiv preprint arXiv:1608.04644*, 2016.

[2] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[3] 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.

[4] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on*, pages 372–387. IEEE, 2016.

[5] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.