Adversarial Attacks and Defenses in Images, Graphs and Text: A Review

- 2020, 108 citations

I hope I can turn these notes into the *Preliminaries* thesis section

Vulnerable networks

- CNN
- FC DNN
- RNN
- GCN (graph convolutional networks) - used in fraud detection
  - only necessary to change couple of edges

Counter-measures

- Gradient masking
- Robust optimization
- Adversary detection

Deep neural-nets reason differntly -> understanding adversarial attacks should help understand this difference

Definitions and notations

Threat model

Adversary's goal

- Poisoning attack - change the behavior of DNN by modifying/inserting few train examples
  - public honeypot - collection of training data for malware detectors
- evasion attack - craft fake examples classifier cannot recognize
  - targeted
  - untargeted

Adversary's knowledge

- White-box attack - widely studied, easily analyzed mathematically
- Black-box attack - practical
- Semi-white (gray) box attack - train generative model in white-box setting, then use in black-box scenario (TREMBA)

Victim models

- conventional machine learning models - SVM, Naive-Bayes
- DNN - not well understood how they work, studying security necessary
  - FC
  - CNN - sparse version of FC
  - GCN
  - RNN

Security evaluation

Robustness

- **Minimal pertubation** - The smallest pertubation to fool the network
  - $\delta_{min} = \arg\min_{\delta} \|\delta\| \text{ s.t. } F(x+\delta) \neq y$
- **Robustness** - The norm of minimal pertubation on particular example
  - $r(x, F) = \|\delta_{min}\|$

- **Global robustness** - Expectation of robustness over the whole dataset
  - $\rho(F) = \mathop{\mathbb{E}}_{x\sim\mathcal{D}} r(x, F)$

Adversarial risk (loss)

- **Most-adversarial example** - Given classifier F, datapoint x and ϵ ball, x_{adv} is the adversarial example with the largest loss
  - x_{adv} is the point, where the classifier is the most likely to be fooled
  - $x_{adv} = \arg\max_{x'} \mathcal{L}(x\ F)$ s.t. $\|x' - x\| \leq \epsilon$
- **Adversarial loss** - Loss value of the most-adversarial example
  - $\mathcal{L}_{adv}(x) = \mathcal{L}(x_{adv}) = \max_{\|x'-x\|<\epsilon} \mathcal{L}(\theta, x', y)$
- **Global adversarial loss (adversarial risk)** - The expectation of adversarial loss over the data distribution $\mathcal{D}$
  - $\mathcal{R}_{adv}(F) = \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \mathcal{L}_{adv}(x) = \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \max_{\|x'-x\|<\epsilon} \mathcal{L}(\theta, x', y)$

Adversarial risk vs. risk

- The concept of *Adversarial risk* is similar to the definition of classifier risk (empirical risk)
  - $\mathcal{R}(F) = \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \mathcal{L}(\theta, x, y)$
  - Global adversarial risk (loss) is in a sense empirical risk but on the most adversarial examples, low empirical risk doesn't have to mean low adversarial risk

Generating adversarial examples

Studying adversarial examples in the image domain essential, because: - perceptual similarity between fake and original is intuitive (unlike in other domains - graphs, audio) - imaga data have simple structure

Studied datasets

- MNIST
- CIFAR10
- ImageNet

All attacks summarization table

White-box attacks

Given classifier C (model F), victim sample (x, y), synthesize fake image x', that is perceputally similar to original x, but fools the classifier C - find x' satisfying $\|x' - x\| \leq \epsilon$, such that $C(x') = t \neq y$ - $\|\cdot\|$ usually l_p norm

Biggio's attack

- ECML 2013
- adversarial examples on MNIST targeting SVMs and 3-layer FC DNNs
- inspired studies on safety of deep learning

Szegedy's limited-memory BFGS (L-BFGS)

- Dec 2013
- first to attack image classifiers
- find minimally distorted adversarial example x' by solving:
  - $\min \|x - x'\|_2^2$, s.t. $C(x') = t$ and $x' \in [0, 1]^m$
- loss function:
  - $\min c\|x - x'\|_2^2 + \mathcal{L}(\theta, x', t)$, s.t. $x' \in [0, 1]^m$
- by increasing constant c throughout the optimization, while keeping the adversarial example outside of the correct decision boundary, we can approximately find adversarial example with minimal pertubation

Fast gradient sign method (FGSM)

- Dec 2014, Goodfellow et al.

- **non-target**
  - $x' = x + \epsilon \text{sgn}(\nabla_x \mathcal{L}(\theta, x, y))$
  - maximise loss of correct classification
- **target**
  - $x' = x - \epsilon \text{sgn}(\nabla_x \mathcal{L}(\theta, x, t))$
  - minimize loss of target class
- For targeted attack setting, this can be framed as one-step 1-bit-precision gradient descent to solve:
  - $\min \mathcal{L}(\theta, x', t)$ s.t. $\|x' - x\|_\infty \leq \epsilon$ and $x' \in [0, 1]^m$
  - resulting x' is vertex (extreme point) of ϵ hypercube around x
- only one backprop, very fast
- used for producing samples for adversarial training

Deep Fool

- Nov 2015
- hyperplane of decision boundary
  - $f(x) = F(x)_y - F(x)_t = 0$
- they linearize this hyperplane using Taylor expansion
  - $f'(x) \approx f(x_0) + \langle \nabla_x f(x_0), (x - x_0) \rangle = 0$
- they find orthogonal vector ω from x_0 to the hyperplane and move along its direction
- for instance LeNet can be fooled on over 90% test samples with $l_\infty \leq 0.1$

Jacobian-based saliency map attack (JSMA)

- Nov 2015, Papernot et al.