A Cryptographic Framework for Machine Learning Security

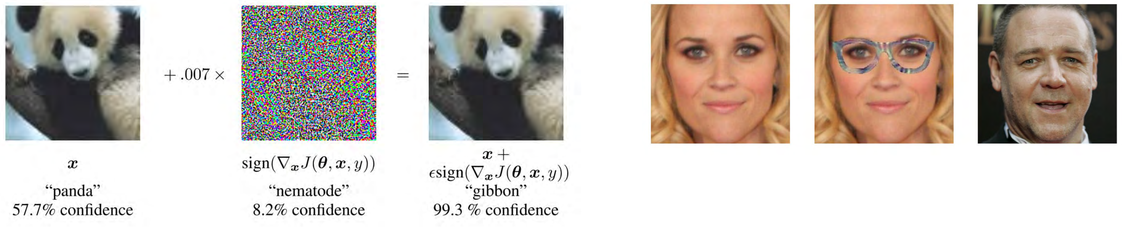
Foundations of Data Science
DATA SCIENCE INSTITUTE

Kevin Shi
Department of Computer Science
Columbia University

Data Science Institute
COLUMBIA UNIVERSITY

The challenge of adversarial perturbations

High dimensional classifiers are not learned perfectly and thus are vulnerable to imperceptible changes which can change the classification output.



Current state-of-the-art defenses train models to be explicitly robust towards specific attacks but are nonetheless broken by different attacks.

Cryptographic approach using hidden random codes

- Allow the model to have non-robust decision boundaries
- Prove computationally intractability for finding and breaking these boundaries

Construction 1 (Random codewords classifier)**.**
Given a multiclass classification problem with data space $(\mathcal{X}, +)$ and labels $\mathcal{Y} = \{1, \cdots, N\}$:

- Sample random code matrix $C \in \{\pm 1\}^{N \times M}$ where $C_{ij} \sim \text{Bernoulli}(1/2)$
- For $j = 1, \cdots, M$, train the binary classifier f_j with $\{i \,|\, C_{ij} = 1\}$ and its complement.
- $F(x) = OO \arg\min_i \|C_i - (f_1(x), \cdots, f_M(x))\|_1$

Table 1: A distributed code for the digit recognition task.

	Code Word					
Class	vl	hl	dl	cc	ol	or
0	0	0	0	1	0	0
1	1	0	0	0	0	0
2	0	1	1	0	1	0
3	0	0	0	0	1	0
4	1	1	0	0	0	0
5	1	1	0	0	1	0
6	0	0	1	1	0	1
7	0	0	1	0	0	0
8	0	0	0	1	0	0
9	0	0	1	1	0	0

Cryptographic security definitions

We propose the following distinction between an adversarial example and a true one:

- A legitimate perturbation affects all data points (e.g. blending in a real image)
- An adversarial perturbation is designed to only affect some target data point

Definition 2 (Security challenge for untargeted attack)**.**
Given a data point (x, y) where $F(x) = y$, construct a perturbation ρ such that

1. $F(x + \rho) \neq F(x)$
2. $\forall \widehat{y} \neq y, \forall \widehat{x}\, s.t.\, F(\widehat{x}) = \widehat{y}, F(\widehat{x} + \rho) = F(\widehat{x})$

Definition 3 (Binary classifier perturbation oracle)**.**
Given a binary classification problem specified by $\phi : \{1, \cdots, N\} \to \{-1, 1\}$ and a class $b \in \{-1, 1\}$, the oracle produces a perturbation ρ such that

1. For any binary classifier $f_\phi : \mathcal{X} \to \{0, 1\}$ trained with labels produced by ϕ, for all x such that $f(x) = b, f(x + \rho) \neq b$
2. For all x, the perturbation ρ is uncorrelated with $\nabla h_{\widehat{\phi}}(x)$ over a randomly sampled $\widehat{\phi}$:

$$\mathcal{N}(0, \sigma^2) \overset{d}{\approx} \left[h_{\widehat{\phi}}(x + \rho) - h_{\widehat{\phi}}(x)\right] \text{ where } \widehat{\phi} \overset{\text{unif}}{\sim} \{-1, 1\}^N$$

Definition 4 (Generic model game)**.**
The adversary may perform the following three operations to try to break the security challenge:

1. Fix a binary classification problem ϕ and query the perturbation oracle for some ρ that breaks any binary classifier for this as per Definition 3
2. Compute the sum of two perturbations $\rho_1 + \rho_2$
3. Check whether a perturbation ρ satisfies the definitions of the security challenge as per Definition 2. Receives only a binary True or False answer.

Future work

- The perturbation oracle intuitively captures all substitute model transfer learning attacks, but is this general enough?
- Empirical verification of the labels being sufficiently uncorrelated is needed.

Acknowledgments

The author would like to thank his advisors Allison Bishop and Daniel Hsu for valuable discussions.

References

Goodfellow, Shlens, Szegedy. *Explaining and Harnessing Adversarial Examples*
Sharif, Bhagavatula, Bauer, Reiter. *Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition*
Dietterich, Bakiri. *Solving Multiclass Learning Problems via Error-Correcting Output Codes*