Adversarial Examples in Random Neural Networks with General Activations

Andrea Montanari and Yuchen Wu

Stanford University

Abstract

Recent theoretical work [1, 2] proved that adversarial examples are ubiquitous in two-layers networks with sub-exponential width and ReLU or smooth activations, and multi-layer ReLU networks with sub-exponential width. We present a result of the same type, with no restriction on width and for general locally Lipschitz continuous activations.

Adversarial Examples

The output of a neural network at test time can be significantly changed by an imperceptible but carefully chosen perturbation of its input. Such perturbed inputs are referred to as **adversarial examples**.

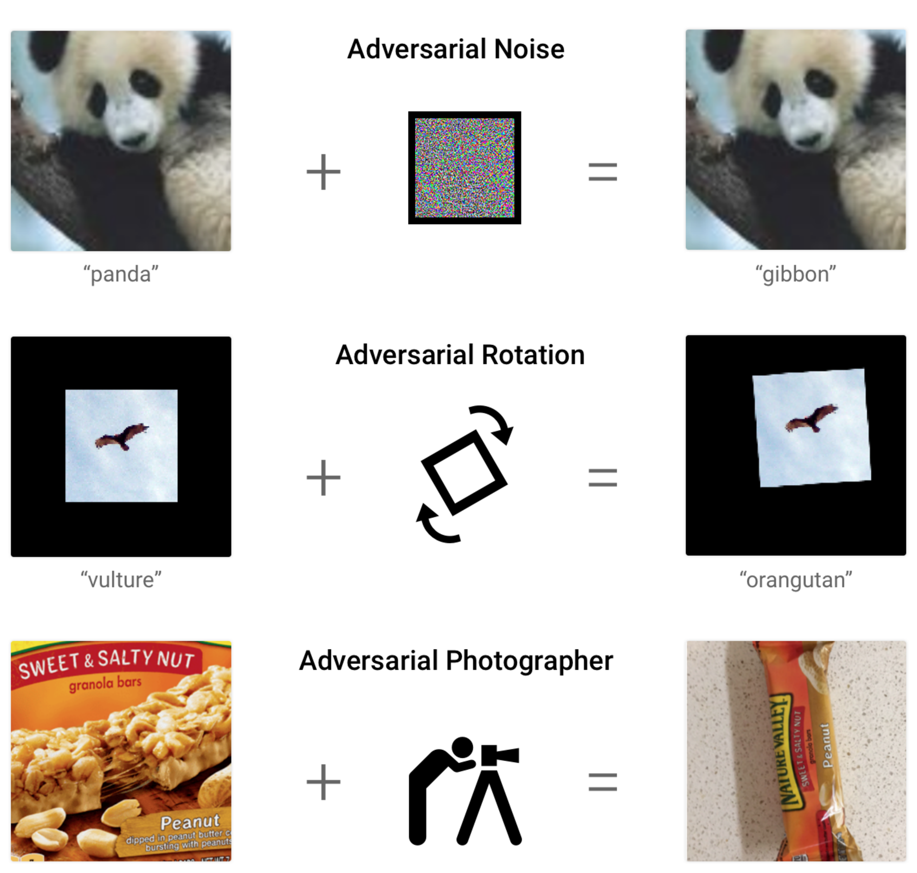

Figure 1: Adversarial examples in real applications.

Assume data sample takes the form $(\boldsymbol{x}, y)$, with $\boldsymbol{x} \in \mathbb{R}^d$ a covariates vector and $y \in \mathbb{R}$ the corresponding label. A model is a function $f(\,\cdot\,;\boldsymbol{\theta}) : \mathbb{R}^d \to \mathbb{R}$ parametrized by weights $\boldsymbol{\theta} \in \mathbb{R}^p$. Given a test point $\boldsymbol{x} \in \mathbb{R}^d$, an adversary constructs $\boldsymbol{x}^s = \boldsymbol{x}^s(\boldsymbol{x};\boldsymbol{\theta}) \in \mathbb{R}^d$. The adversary is successful if, with high probability

$$\operatorname{sign}(f(\boldsymbol{x}^s;\boldsymbol{\theta})) = -\operatorname{sign}(f(\boldsymbol{x};\boldsymbol{\theta})), \tag{1}$$

$$\|\boldsymbol{x}^s - \boldsymbol{x}\| \ll \|\boldsymbol{x}\|. \tag{2}$$

Fast Gradient Sign Method

The fast gradient sign method (FGSM) is an efficient algorithm used to find adversarial examples. More precisely, FGSM can be stated as follows:

$$\boldsymbol{x}^s := \boldsymbol{x} - \tau s_d \nabla f(\boldsymbol{x}),$$

where $\tau := \operatorname{sign}(f(\boldsymbol{x}))$, and $s_d \in \mathbb{R}^+$ is the step size.

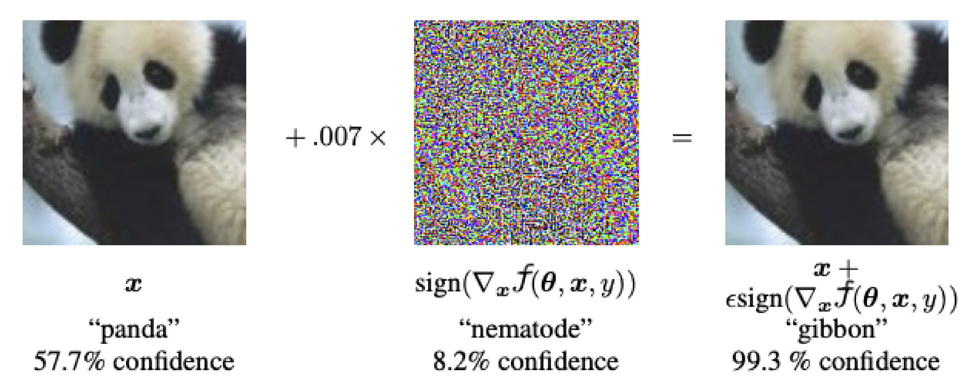

Figure 2: Illustration of fast gradient sign method.

Random Multi-layer Networks

We consider a multi-layer neural network with $l+1$ layers for $l \in \mathbb{N}_+$:

$$f(\boldsymbol{x}) = \boldsymbol{W}_{l+1}\sigma(\boldsymbol{W}_l\sigma(\cdots\sigma(\boldsymbol{W}_2\sigma(\boldsymbol{W}_1\boldsymbol{x}))\cdots)).$$

- $\boldsymbol{W}_i \in \mathbb{R}^{d_i \times d_{i-1}}$
- $(\boldsymbol{W}_i)_{jj'} \overset{iid}{\sim} \mathsf{N}(0, 1/d_{i-1})$ for all $j \in [d_i], j' \in [d_{i-1}]$
- $\{\boldsymbol{W}_i\}_{i \in [l+1]}$ are independent of each other
- Assume $d_0 = d$, $d_{l+1} = 1$, and $d_i = d_i(d) \to \infty$ for all $0 \le i \le l$
- Activation function $\sigma : \mathbb{R} \to \mathbb{R}$ is understood to act on vectors entrywise

Main Result (Informal)

FGSM-like attack finds adversarial examples for neural networks with random Gaussian weights. Comparing to earlier works, our results apply to arbitrary diverging width and general activation functions.

Main Theorem

Let $\boldsymbol{x} \in \mathbb{R}^d$ be a deterministic vector with $\|\boldsymbol{x}\|_2 = \sqrt{d}$. Assume that σ is (1) not a constant, (2) continuous, (3) almost everywhere differentiable, (4) σ' is almost everywhere continuous and (5) σ' is pseudo-Lipschitz. Then the following hold:

Theorem 2.1 of [3]

Let $\{\xi_d\}_{d \in \mathbb{N}_+} \subseteq \mathbb{R}^+$ be an increasing sequence such that $\xi_d \to \infty$ as $d \to \infty$. Then there exists $\{s_d\}_{d \in \mathbb{N}_+} \subseteq \mathbb{R}^+$, such that $s_d \le \xi_d$ and the following hold:

- $\text{p-lim}_{m,d\to\infty} \frac{\|\boldsymbol{x}-\boldsymbol{x}^s\|_2}{\|\boldsymbol{x}\|_2} = 0$,
- $\lim_{m,d\to\infty} \mathbb{P}(\operatorname{sign}(f(\boldsymbol{x})) \neq \operatorname{sign}(f(\boldsymbol{x}^s))) = 1$.

Gaussian Conditioning Lemma

Let $\boldsymbol{X} \in \mathbb{R}^{m \times d}$ have i.i.d. standard Gaussian entries, $\boldsymbol{a} \in \mathbb{R}^d$, $\boldsymbol{w} \in \mathbb{R}^m$. Furthermore, $\boldsymbol{X}$, $\boldsymbol{a}$ and $\boldsymbol{w}$ are mutually independent

- $\boldsymbol{g} = \boldsymbol{W}\boldsymbol{a}$
- $\boldsymbol{m} = \boldsymbol{W}^\intercal F(\boldsymbol{g}, \boldsymbol{w})$, $F : \mathbb{R}^{2m} \to \mathbb{R}^m$

Conditioning on $(\boldsymbol{g}, \boldsymbol{m})$, the conditional distribution $\boldsymbol{X} \mid \boldsymbol{g}, \boldsymbol{m}$ is equal to

$$\Pi_{\boldsymbol{x}}^{\perp}\tilde{\boldsymbol{X}}\Pi_{\boldsymbol{g}}^{\perp} + \Pi_{\boldsymbol{x}}^{\perp}\boldsymbol{X}\Pi_{\boldsymbol{g}} + \Pi_{\boldsymbol{x}}\boldsymbol{X}\Pi_{\boldsymbol{g}}^{\perp} + \Pi_{\boldsymbol{x}}\boldsymbol{X}\Pi_{\boldsymbol{g}},$$

- $\Pi_{\boldsymbol{y}}$ denotes projection onto the subspace spanned by $\boldsymbol{y}$
- $\Pi_{\boldsymbol{y}}^{\perp} = \mathsf{I} - \Pi_{\boldsymbol{y}}$
- $\tilde{\boldsymbol{X}}$ is an independent copy of $\boldsymbol{X}$ that is independent of $(\boldsymbol{g}, \boldsymbol{m})$

Conclusion

Fully connected neural networks with constant depth, Gaussian weights, general activation function and arbitrary diverging width have adversarial examples that can be found by FGSM.

Open Problems

- Diverging depth
- Beyond Gaussian weights
- More complicated structure

References

[1] Sébastien Bubeck, Yeshwanth Cherapanamjeri, Gauthier Gidel, and Remi Tachet des Combes.
A single gradient step finds adversarial examples on random two-layers neural networks.
Advances in Neural Information Processing Systems, 34:10081–10091, 2021.

[2] Peter Bartlett, Sébastien Bubeck, and Yeshwanth Cherapanamjeri.
Adversarial examples in multi-layer random relu networks.
Advances in Neural Information Processing Systems, 34:9241–9252, 2021.

[3] Andrea Montanari and Yuchen Wu.
Adversarial examples in random neural networks with general activations.
arXiv preprint arXiv:2203.17209, 2022.

Acknowledgements

This work was supported by the NSF through award DMS-2031883, the Simons Foundation through Award 814639 for the Collaboration on the Theoretical Foundations of Deep Learning, the NSF grant CCF-2006489, the ONR grant N00014-18-1-2729.

Contact Information

- Web: https://wuyc0114.github.io./
- Email: wuyc14@stanford.edu
- Phone: +1 (650) 613 8374