A History of Adversarial Attacks

Prof. Stephen Roberts[1]; Benjamin Etheridge[1]

AI Security Reading Group, October 2024

[1]Machine Learning Research Group
Department of Engineering Science

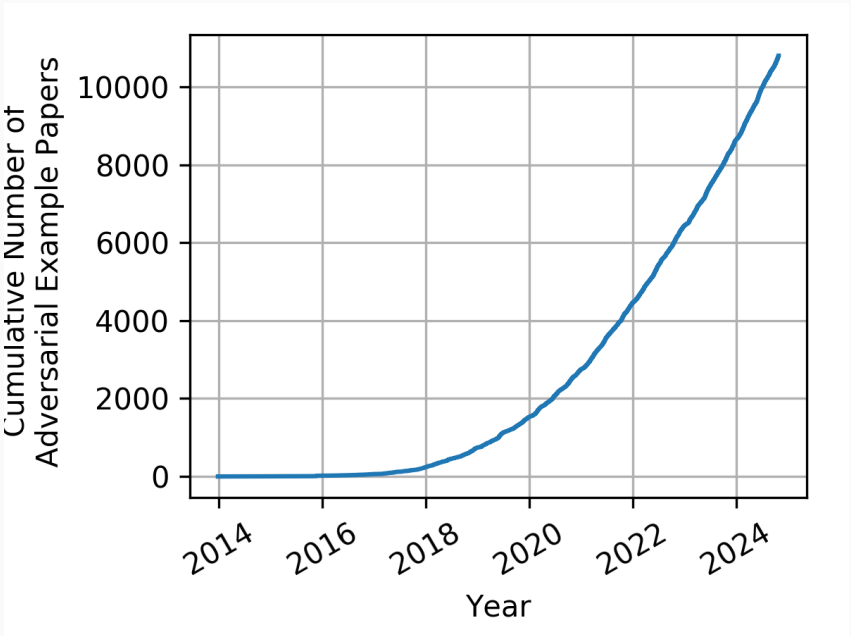
Table of Contents

Table of Contents

Intro - Why should we care?

Intro - Why should we care?

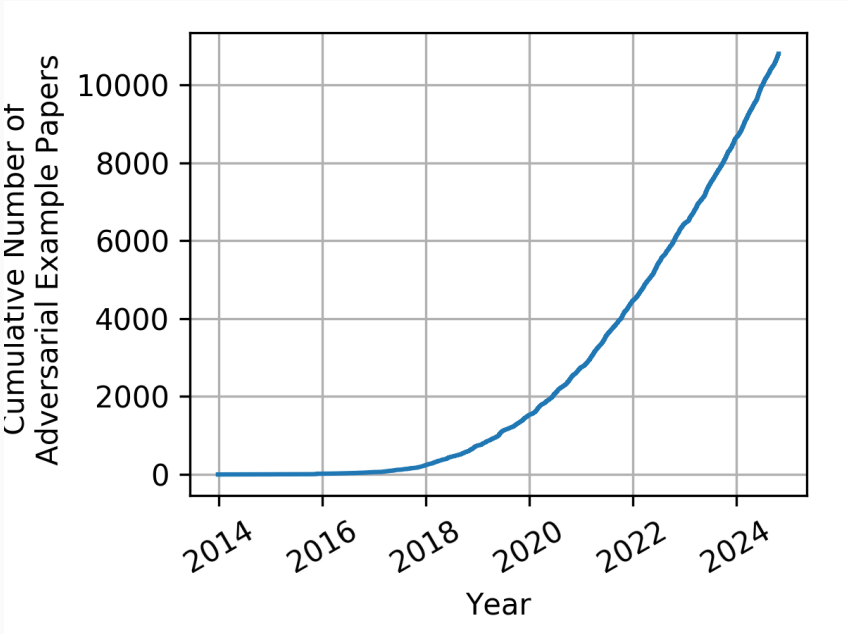
- Extremely active field



https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html

Intro - Why should we care?

- Extremely active field



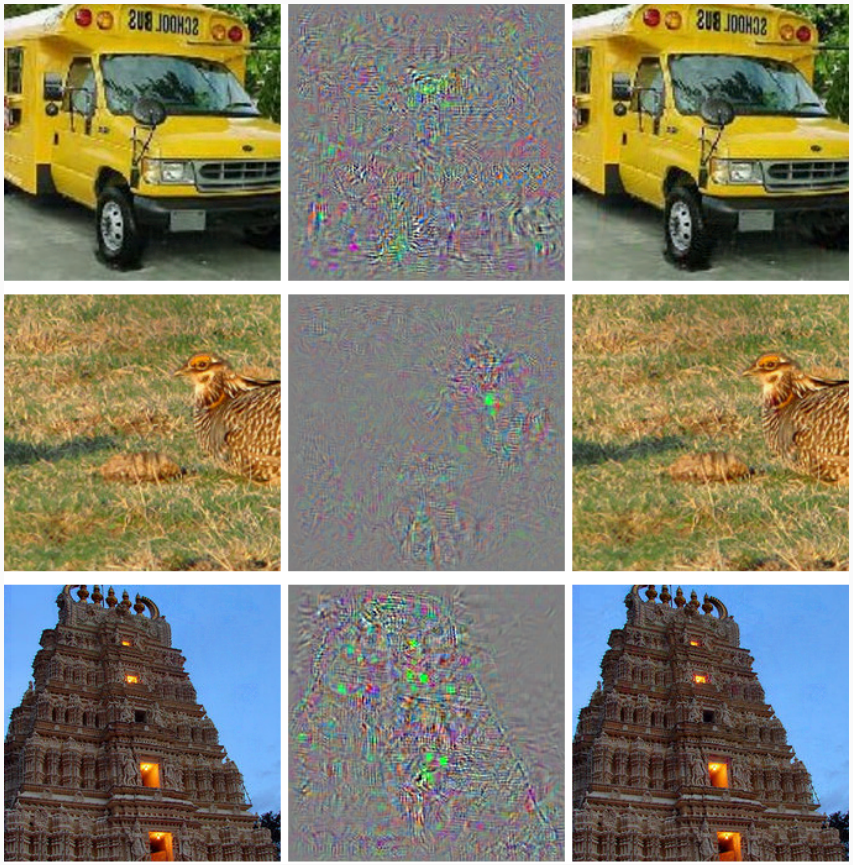
https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html

- Threat model has changed - no longer purely an academic concern.

Intriguing properties of neural networks

Szegedy et al. - ICLR 2014 - arXiv:1312.6199

- Imperceptible perturbations can result in significant changes in output



Intriguing properties of neural networks

Szegedy et al. - ICLR 2014 - arXiv:1312.6199

- Box constrained L-BFGS Computation, minimising $||L||_2$ of perturbation η s.t:
  1. $f(x + \eta) = l$ (misclassification to label l)
  2. $x + \eta \in [0, 1]^m$ (?)

Intriguing properties of neural networks

Szegedy et al. - ICLR 2014 - arXiv:1312.6199

- Box constrained L-BFGS Computation, minimising $||L||_2$ of perturbation η s.t:
  1. $f(x + \eta) = l$ (misclassification to label l)
  2. $x + \eta \in [0, 1]^m$ (?)
- Second Order Optimisation - Inefficient

Intriguing properties of neural networks

Szegedy et al. - ICLR 2014 - arXiv:1312.6199

- Box constrained L-BFGS Computation, minimising $||L||_2$ of perturbation η s.t:
  1. $f(x + \eta) = l$ (misclassification to label l)
  2. $x + \eta \in [0, 1]^m$ (?)
- Second Order Optimisation - Inefficient
- ICLR 2024 Test Of Time Runner Up

Explaining and Harnessing Adversarial Examples

Goodfellow et al. - ICLR 2015 - arXiv:arXiv:1412.657

- Introduces Fast Gradient Sign Method (FGSM)

TITLE

AUTHOR1 et al. YEAR - arXiv:REF

-

TITLE

AUTHOR1 et al. YEAR - arXiv:REF

TITLE

AUTHOR1 et al. YEAR - arXiv:REF

TITLE

AUTHOR1 et al. YEAR - arXiv:REF

Table of Contents

Sample frame title

In this slide, some important text will be highlighted because it's important. Please, don't abuse it.

Remark
Sample text

Important theorem
Sample text in red box

Examples
Sample text in green box. The title of the block is "Examples".

Two-column slide

This is a text in first column.

$$E = mc^2$$

- First item
- Second item

This text will be in the second column and on a second tought this is a nice looking layout in some cases.