PyData Delhi 2018

Fooling and Protecting Deep learning models

Divyam Madaan

Twitter: @divyam3897
Github: @divyam3897

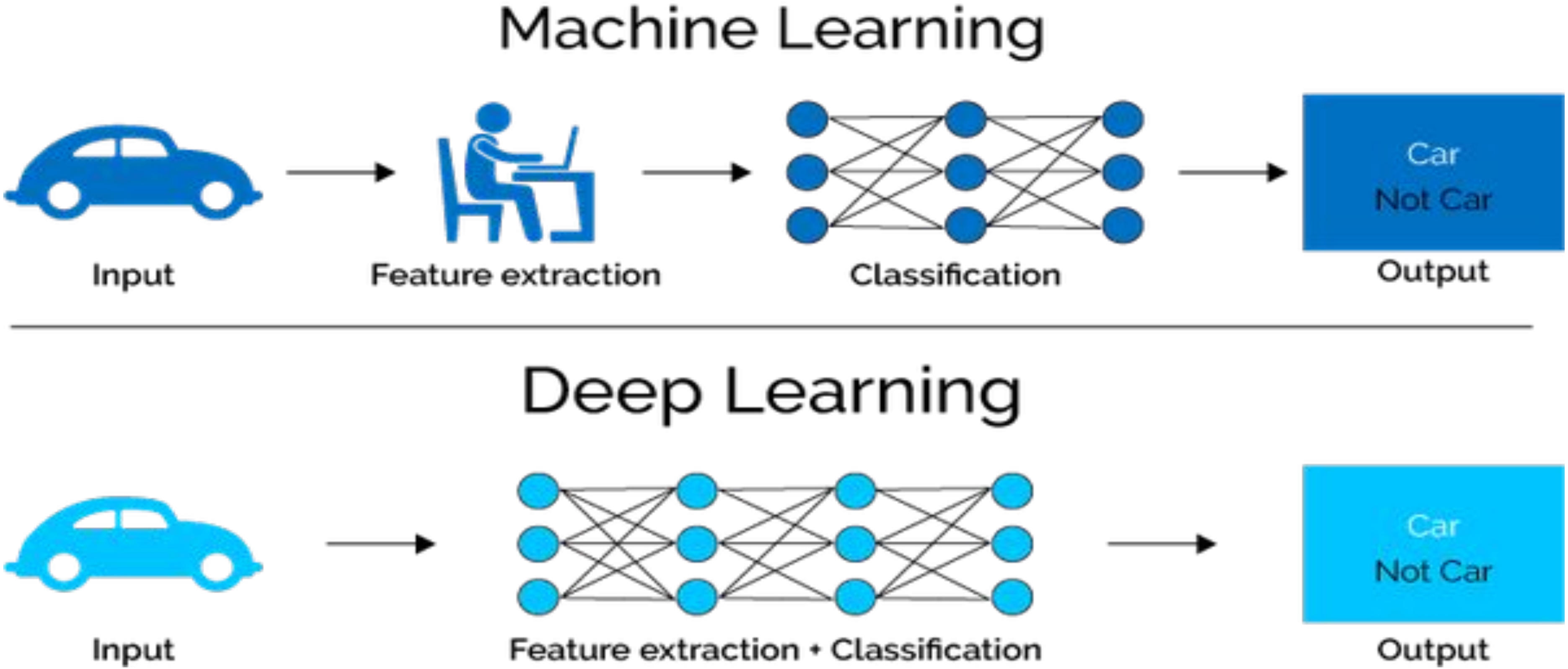
About me

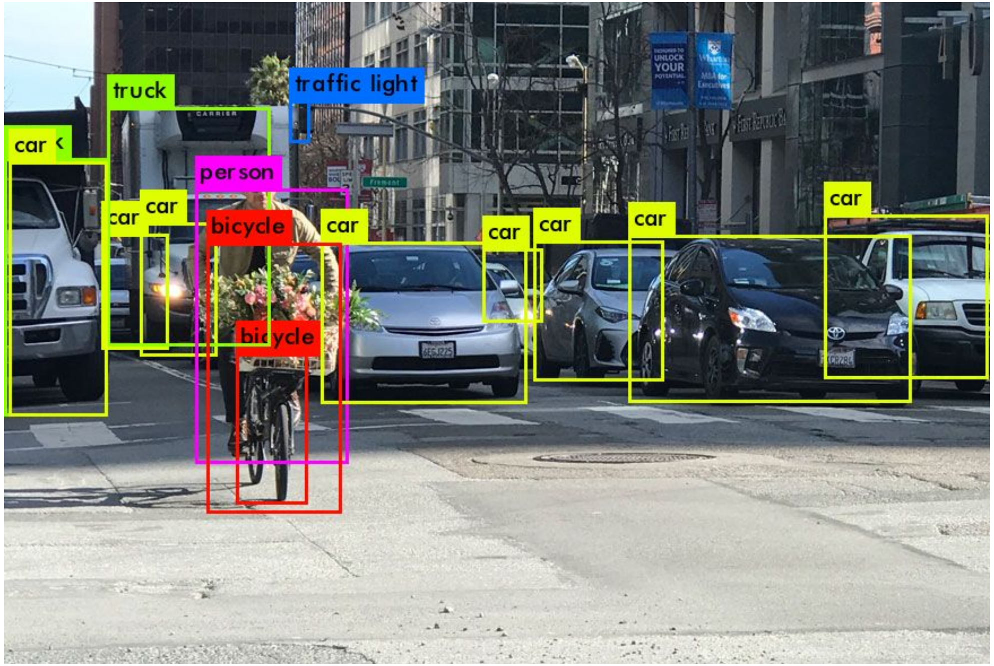
- Final year undergraduate
- Open source enthusiast
- Deep Learning Researcher at FOR.ai

Agenda

- ➔ Machine learning and Deep learning overview
- ➔ Attacks on machine learning pipeline
- ➔ Types of attacks
- ➔ Timeline of machine learning security
- ➔ Defending the machines

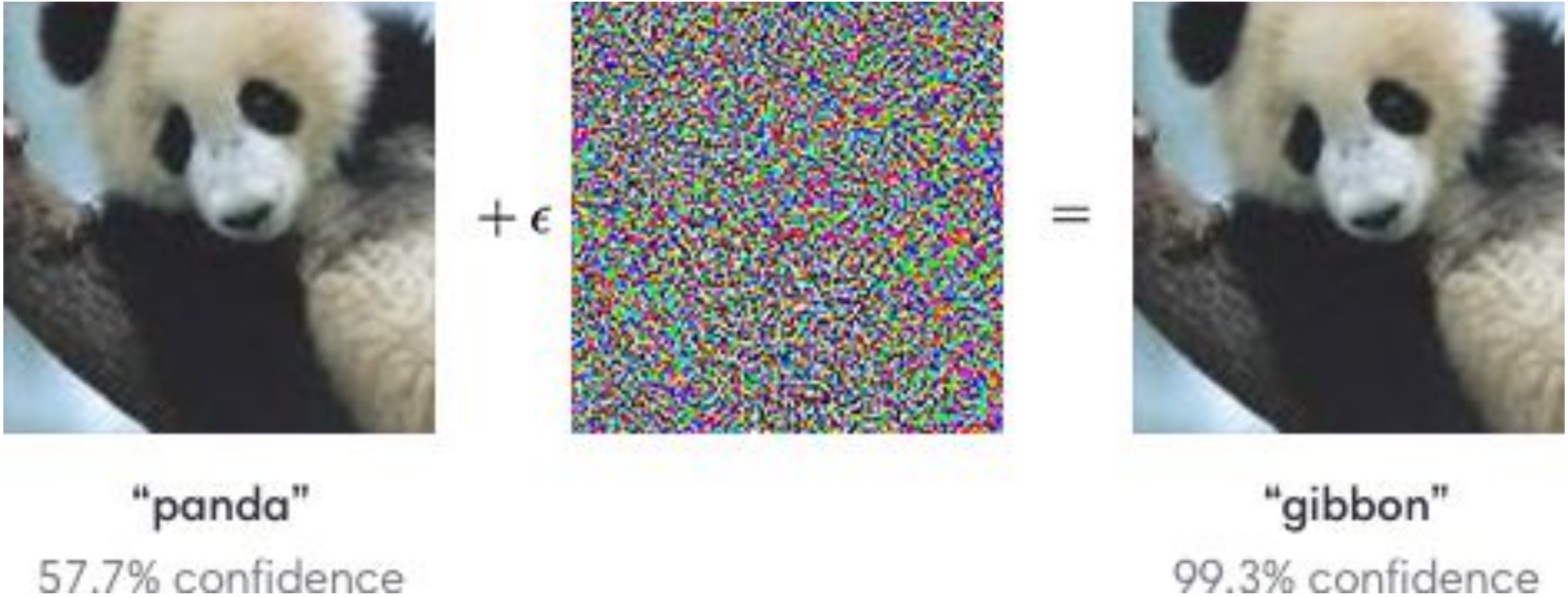
Machine learning vs Deep learning



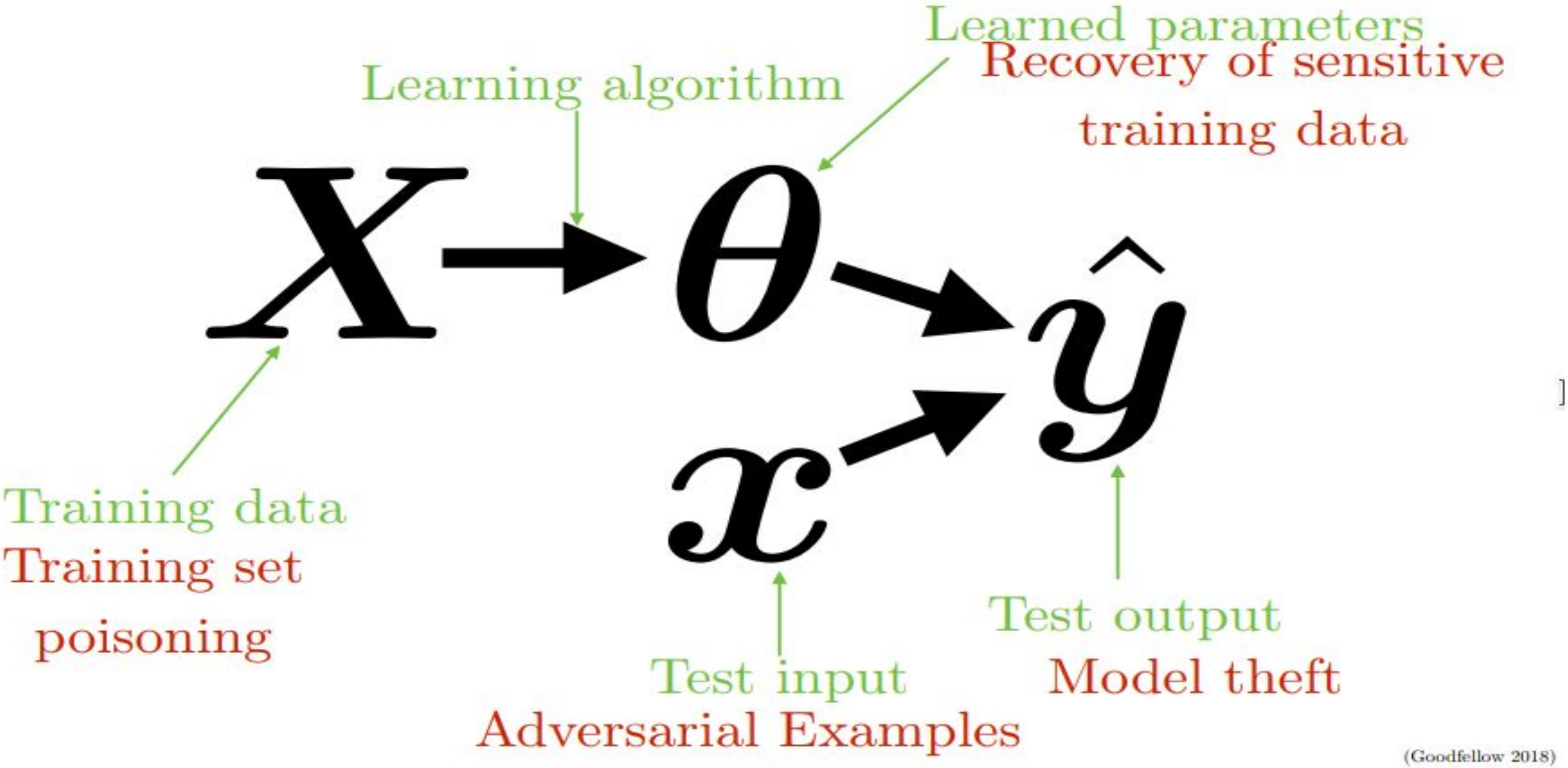


Reached state of the art on various tasks

Good models make mistakes



Attack on the machine learning pipeline



Types of Attacks

White box Attacks

Black box attacks

Targeted Attacks

Untargeted Attacks

Not just images!

Deep Text Classification Can be Fooled

Bin Liang and **Hongcheng Li** and **Miaoqiang Su** and **Pan Bian** and **Xirong Li** and **Wenchang Shi**
School of Information, Renmin University of China, Beijing, China
{liangb, owenlee, sumiaoqiang, bianpan, xirong, wenchang}@ruc.edu.cn

Did you hear that? Adversarial Examples Against Automatic Speech Recognition

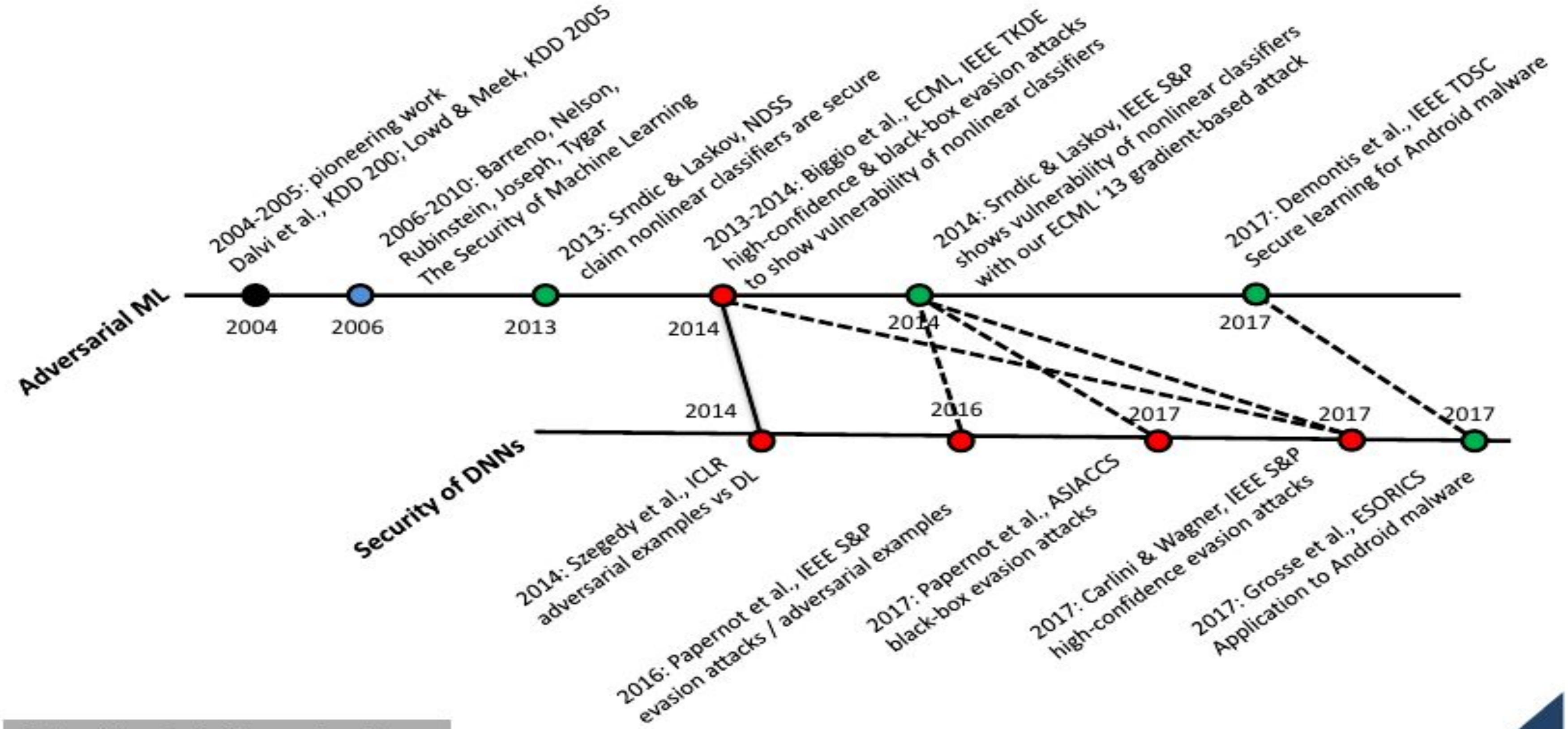
Robust Physical Adversarial Attack on Faster R-CNN Object Detector

Shang-Tse Chen[1], Cory Cornelius[2], Jason Martin[2], and Duen Horng (Polo) Chau[1]

Face Recognition on Consumer Devices: Reflections on Replay Attacks

Daniel F. Smith, Arnold Wiliem *Member, IEEE*, and Brian C. Lovell *Senior Member, IEEE*

Timeline of Learning Security



What this means for us?

- Deep learning algorithms (Machine learning in general) are susceptible to attacks.
- Use with caution in critical deployments
- Spend effort to make model robust to tempering
- Evaluate a model’s adversarial resilience - not just accuracy/precision/recal.

Defending the machines

Pre-2013: Defenses for convex models

Szegedy et al 2013: train on adversarial examples

Goodfellow et al 2014: generate them constantly in the inner loop of training (minimax)

Kurakin et al 2016: use an iterative attack

Madry et al 2017: randomize the starting point of the attack. 1st to generalize over attack algorithms

Kannan et al 2018: logit pairing

(Goodfellow 2018)


LIVE DEMO

Resources

- ➔ Breaking Linear classifiers with convnets - Andrej Karpathy
- ➔ Attacking machine learning with adversarial examples - Open AI
- ➔ Adversarial Examples and adversarial training - Ian Goodfellow
- ➔ Adversarial examples in machine learning - Ian Goodfellow
- ➔ ICCV Tutorial on machine learning - Battista Biggio and Fabio Roli
- ➔ Cleverhans

Thank you
Questions?

Twitter: @divyam3897

Linkedin: https://www.linkedin.com/in/dmadaan3897

Github: @divyam3897