Y-DATA 2nd Research Seminar

2025

Explaining and harnessing adversarial examples

Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy (2015)

By Yair and Kai

Adversarial Machine Learning: Impact of Goodfellow et al. (2015)

Guideline Interests:

- Were the authors the first to create adversarial examples
- Subsequent works
- What impact has this paper upon the field
- Criticism

Part 1

Were the authors the first to create adversarial examples

Origins and Early Methods of Adversarial Examples

First Discoveries

- Adversarial classification Dalvi et al. (2004)
- Feature cross-substitution in adversarial classification *Li and Vorobeychik 2014 (neurips)*

Preceding works

- Evasion attacks against machine learning *Biggio et al. (2013)*

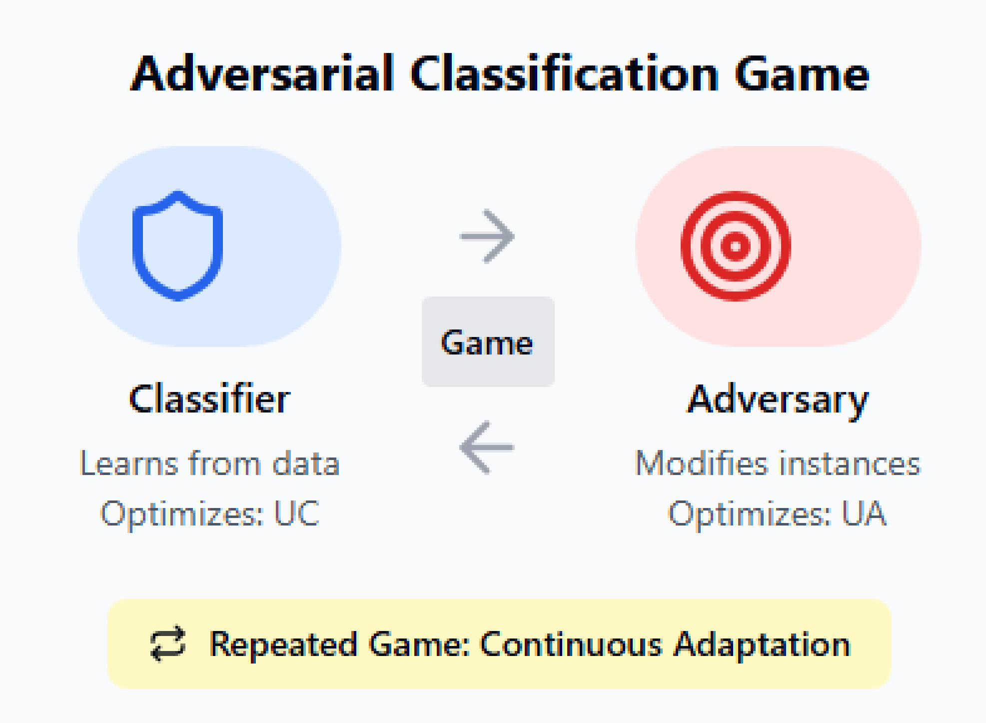
Early Generation Methods

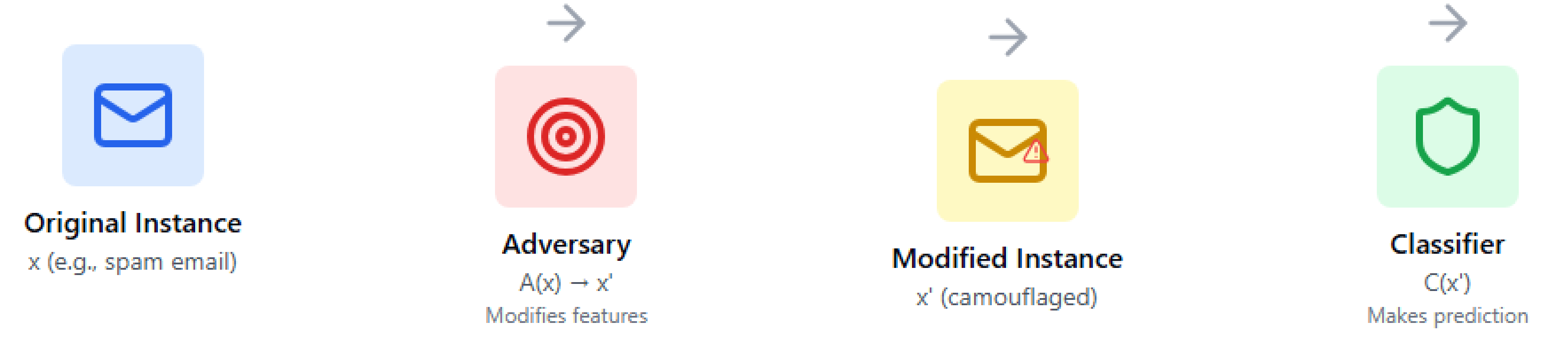
- Intriguing properties of neural networks *Szegedy et al. (2014)*

Origins and Early Methods of Adversarial Examples

1

Dalvi et al. (2004) - "Adversarial Classification"





Traditional Naive Bayes

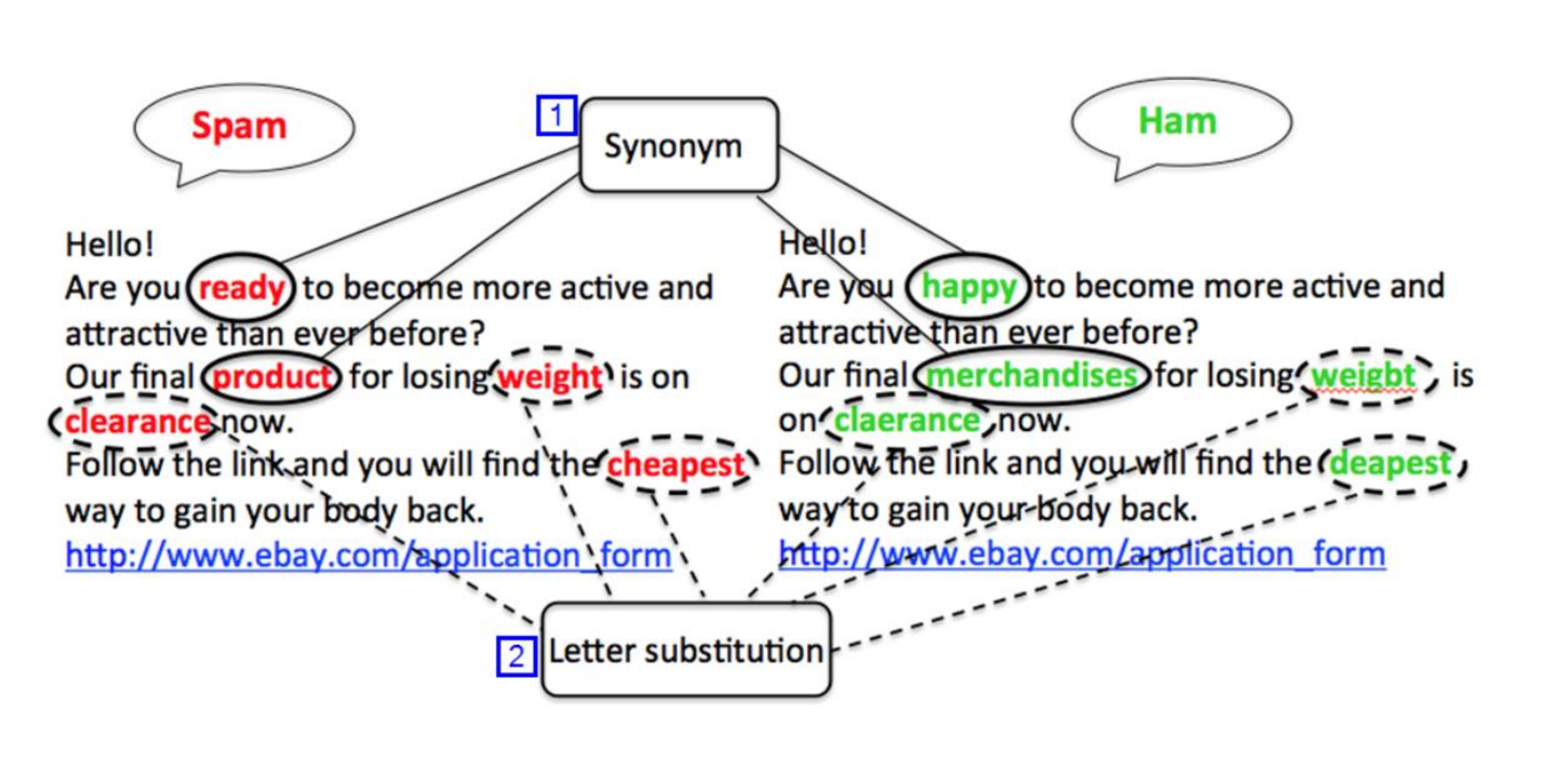
- Assumes data is untainted
- Performance degrades with adversarial manipulation
- High false negative rate (misses spam)

Adversarial Classifier

- Accounts for optimal adversary strategy
- Maintains performance under attack
- Minimizes false positives

Origins and Early Methods of Adversarial Examples

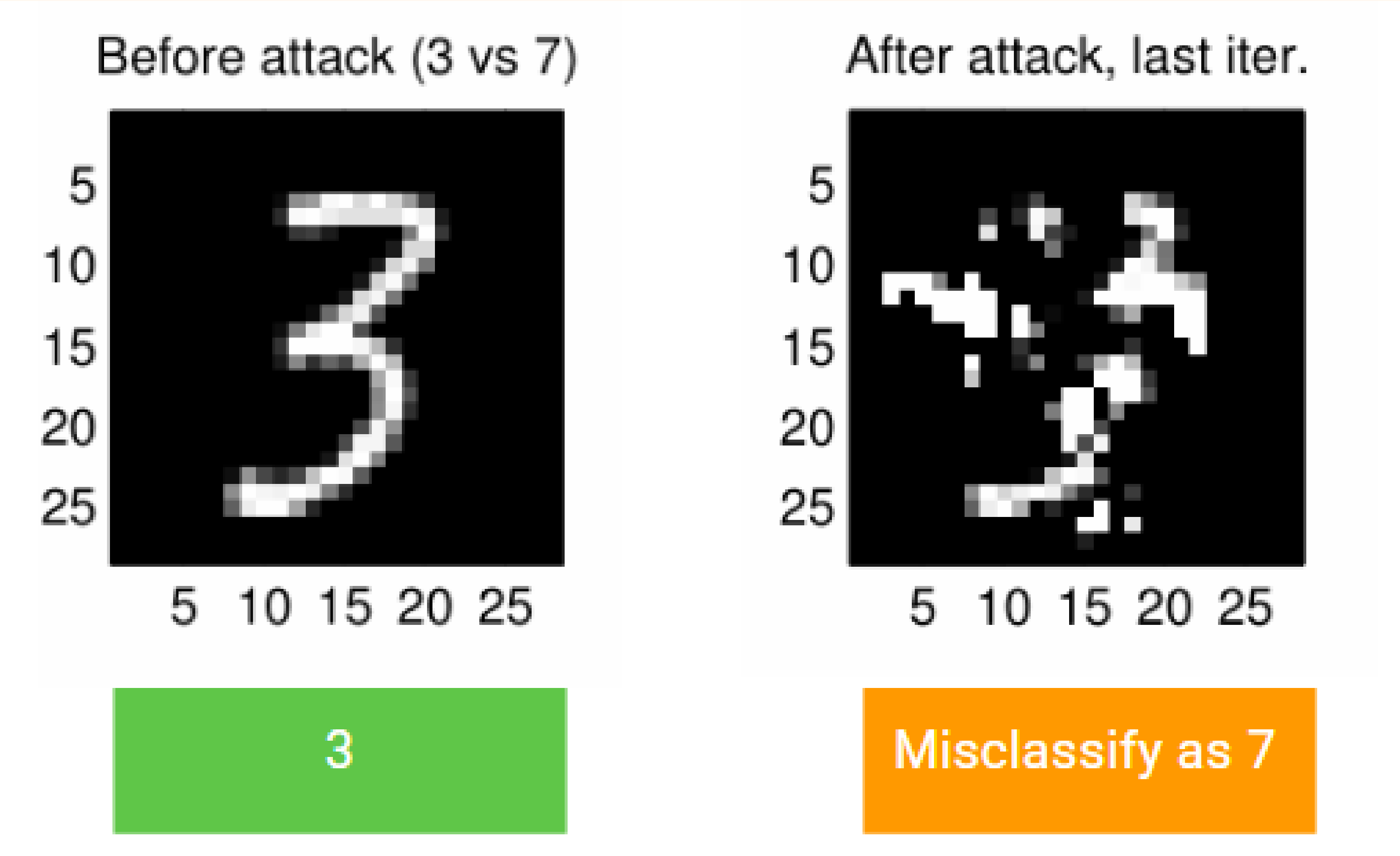
2



Feature cross-substitution in adversarial classification Li and Vorobeychik 2014 (NEURIPS)

Origins and Early Methods of Adversarial Examples

3



Evasion attacks against machine learning Biggio et al. (2013)

Part 2

What impact has this paper upon the field

Academic and Industry Impact of Goodfellow et al.

Fundamental Understanding

Clarified adversarial vulnerability arises from neural networks' linearity in high-dimensional spaces.

- 10,000+ citations
- Created field: Adversarial ML

1

Methodological Advances

FGSM enabled practical adversarial training and widespread research on attacks and defenses.

- 1000x faster than L-BFGS
- Spawned 10+ attack methods

2

Security Awareness

Highlighted risks in deploying ML systems, spurring robust AI development.

- Google, Meta, Microsoft teams
- New field: ML Security

3

Theoretical Insights

Revealed trade-offs between optimization ease and robustness, inspiring new regularization approaches.

- New conference tracks
- Robustness-accuracy trade-off

4

2015	2017	2019	2021	2023
FGSM Paper	Physical Attacks	Certified Defenses	Benchmarks	LLM Attacks

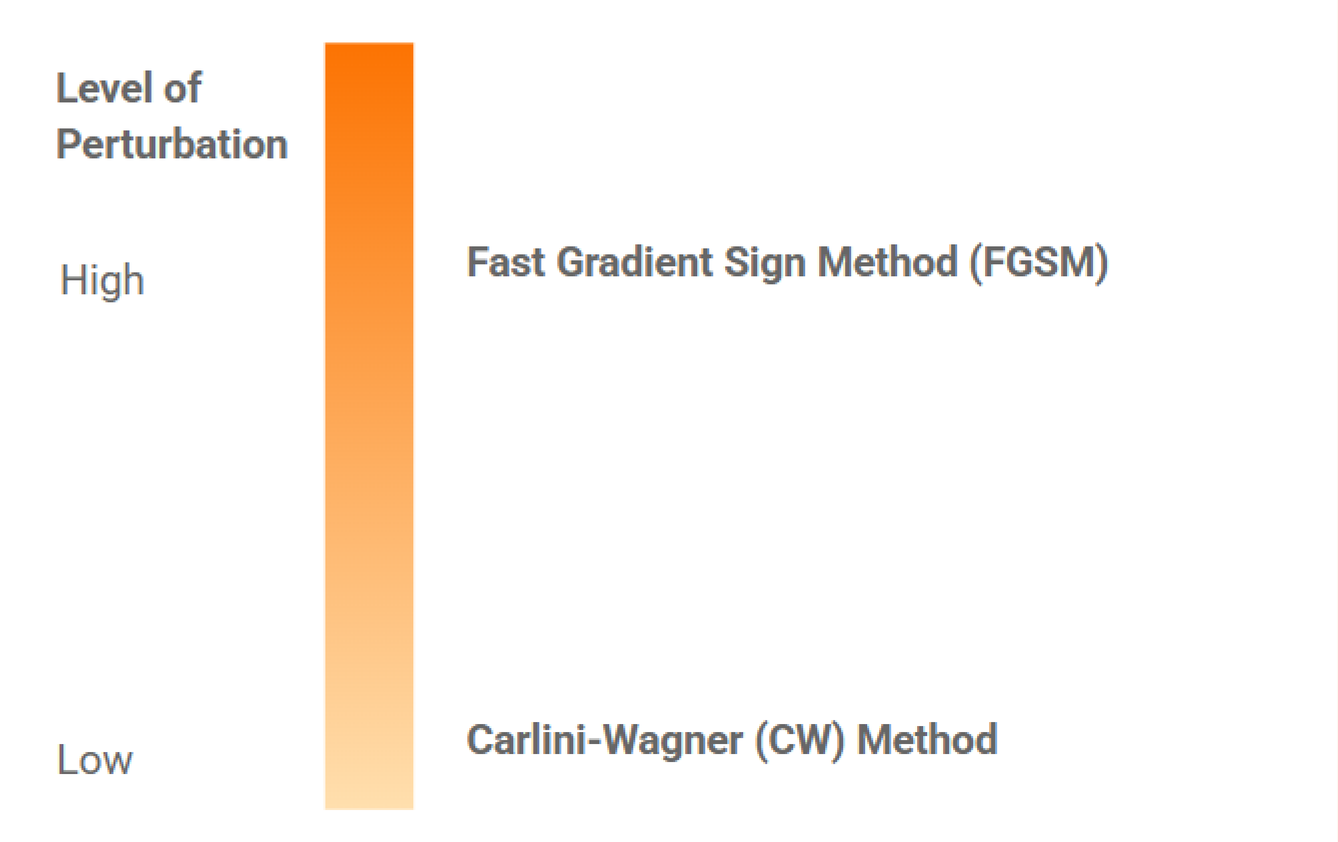
Kurakin, Goodfellow & Bengio - "Adversarial examples in the physical world"
"Certified Adversarial Robustness via Randomized Smoothing" (ICML 2019)
Wei et al. - "Jailbroken: How Does LLM Safety Training Fail?" (2023)

Part 3

Subsequent works

What is perturbation?

Amount of deliberate modification to an input,
so that model outputs an incorrect class.



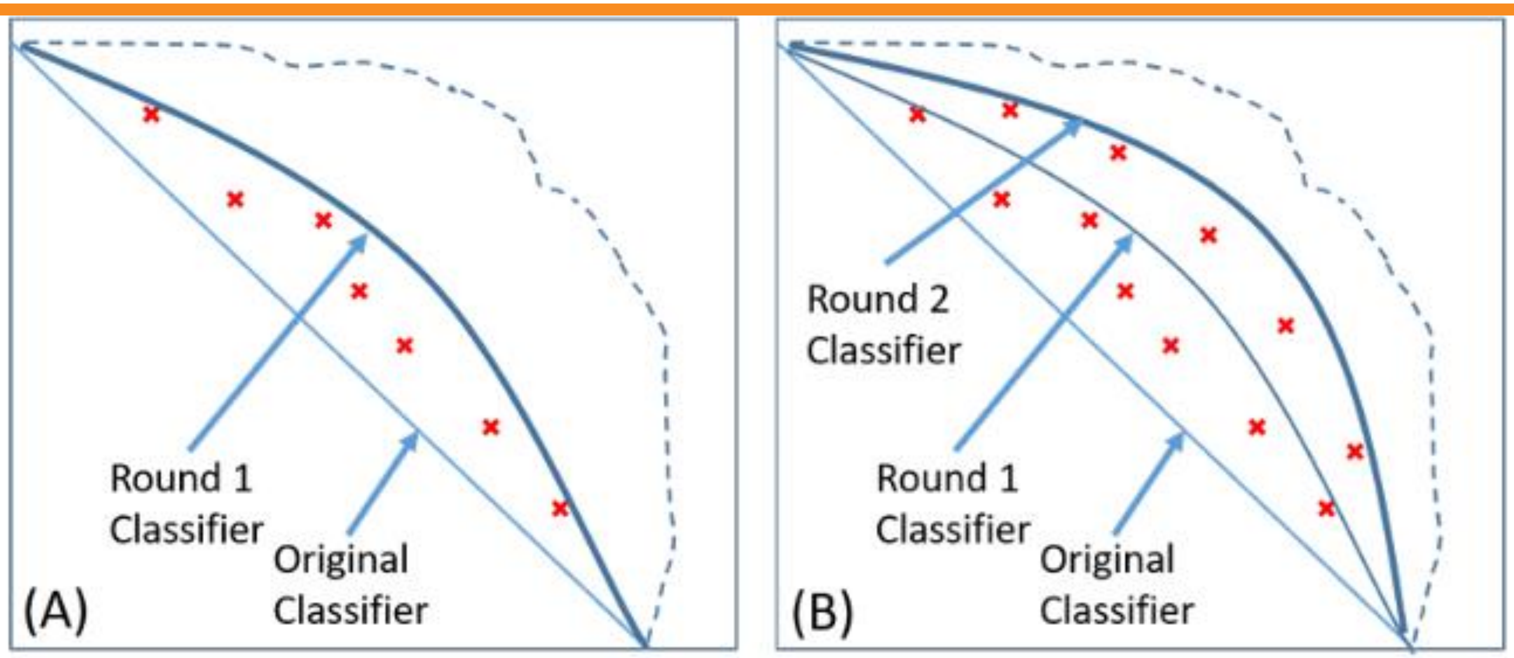
Goodfellow et al (2015)

Carlini & Wagner Attack (2016)

Level of Perturbation

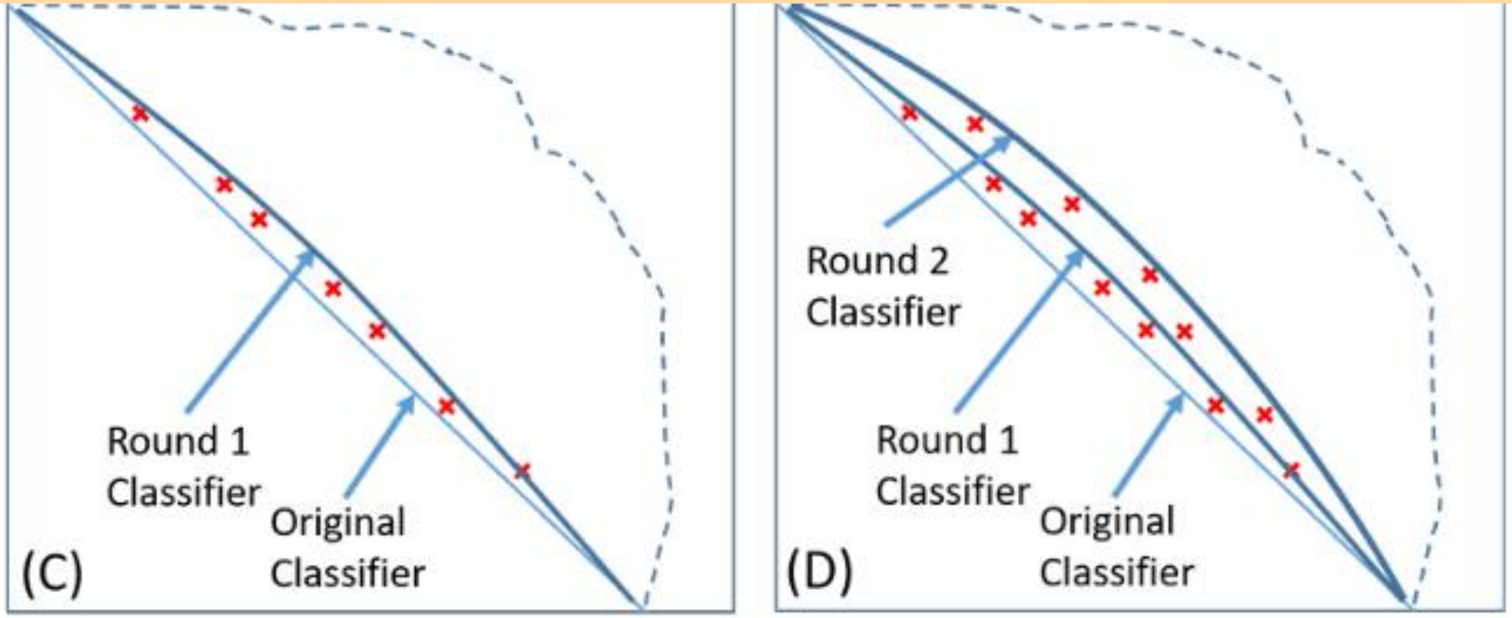
High

Low



DLN for high perturbation attacks (A) is after one round, (B) is after two rounds
Dashed line is true separator; x-marks are adversarial examples from all rounds

Big change



DLN for low perturbation attacks (C) is after one round, (D) is after two rounds
Dashed line is true separator; x-marks are adversarial examples from all rounds

Small change

Alternative Attack Methods Beyond FGSM

DeepFool (2016)

Iteratively finds minimal perturbations crossing decision boundaries, producing smaller changes than FGSM.

Carlini & Wagner Attack (2016)

Low perturbation technique. Highly effective against defenses like defensive distillation, exposing vulnerabilities in many models.

Universal Adversarial Perturbations

Single perturbations fool models across inputs, generated in supervised and unsupervised settings, assessing robustness broadly.

Nguyen et al (2018)

A Learning and Masking Approach to Secure Learning.

Able to **generate a mixture of low and high perturbation examples**

🛡️ Defense Mechanisms Evolution Post-Goodfellow 🛡️

Defense Mechanisms Inspired by Goodfellow et al.

Denoising Autoencoders (DAEs)

Input preprocessing defense

Can remove significant adversarial noise but may not fully secure networks when combined with original models. Uses reconstruction loss to filter perturbations.

Meng & Chen, MagNet (2017)

Robust Architectures

Defense Learning NN

Exploration of network topologies (skip connections, dense nets) and preprocessing to enhance resistance. Lipschitz constraints limit gradient explosions.

Cisse et al., Parseval Networks (2017)

Certified Robustness

Provable guarantees

Mathematical verification that no adversarial example exists within L-p ball. Uses interval bound propagation or convex relaxations for formal guarantees.

Wong & Kolter (2018), Cohen et al. (2019)

Randomized Smoothing

Statistical certification

Adds Gaussian noise during inference to create smooth classifiers with probabilistic guarantees. Trades accuracy for certified radius.

Cohen et al., (ICML 2019)

Feature Denoising & Scattering

Mid-layer defense

Removes adversarial patterns in feature space using wavelet scattering networks or feature statistics. More effective than input denoising alone.

Rauber et al. (2017), Xie et al. (2019)

Ensemble Defenses

Diversity-based protection

Multiple models with different architectures vote on predictions. Adversarial Ensemble Training (AET) trains diverse models jointly against various attacks.

Pang et al., AET (2019)

Timeline of Techniques (Torchattacks Library)

2015	2016	2017	2019		2021	2022
Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2014)	**Carlini-Wagner Method (CW)** (Carlini et al., 2016)	**Projected Gradient Descent (PGD)** (Mardry et al., 2017)	**Fast Adaptive Boundary (FAB)** (Croce et al., 2019)	**OnePixel** (Su et al., 2019) **SparseFool -** Few pixels (Modas et al., 2019)	**Jitter** (Schwinn, et al., 2021)	**Pixle -** Rearrange pixels (Pomponi, et al., 2022)

And variants of

- Fast Gradient Sign Method (FGSM)
- Projected Gradient Descent (PGD)

Sources:
Techniques in the Torchattack library: https://github.com/Harry24k/adversarial-attacks-pytorch/tree/master?tab=readme-ov-file

Notebook Exploring 6 Torchattack Techniques

- **Task**: Image classification
- **Model**: VGG16 with pre-trained weights
- **Dataset**: ImageNet1000 (because VGG16 was trained on this dataset)
- **6 Techniques**:
  (Base) FGSM
  Carlini-Wagner Method (CW)
  (Base) PGD Projected Gradient Descent
  Fast Adaptive Boundary (FAB)
  OnePixel
  Jitter

Original Image
True: confectionery
Predicted: confectionery (90.2%)

Notebook Exploring 6 Torchattack Techniques

Output similar for **Fast Adaptive Boundary (FAB)**

Notebook Exploring 6 Torchattack Techniques

Output similar for **Carlini-Wagner Method (CW), Projected Gradient Descent (PGD), Jitter**

Expanding Attack Knowledge

Broader Attacks

Extended beyond images to speech and NLP, including hidden commands in speech recognition demonstrated by Carlini et al.

Part 4

Criticism

Criticisms and Limitations of the Paper

Oversimplification of Linearity

Critics argue **nonlinearities and complex interactions also contribute** to adversarial vulnerability. (*Ilyas et al. (2019)* .

Adversarial Training Limits

FGSM-based training struggles against **stronger attacks and large datasets, with high computational cost.**

Focus on Toy Datasets

Experiments on MNIST limit **real-world applicability**; physical-world attacks show broader challenges. (Kurakin et al. 2017)