EDGE AI Academy

Summer School

7-8 July 2025, Pisa, Italy

Edge AI Academy Summer School

AI under attack: Navigating the security and privacy frontier

Ihsen Alouani,
CNRS, Université Polytechnique Hd France
Queen's University Belfast

7-8 July 2025 Pisa, Italy

Outline

- AI Background
- Security in AI
- Privacy Risks
- GenAI and Agentic AI Security
- Conclusion

Copyright © 2025

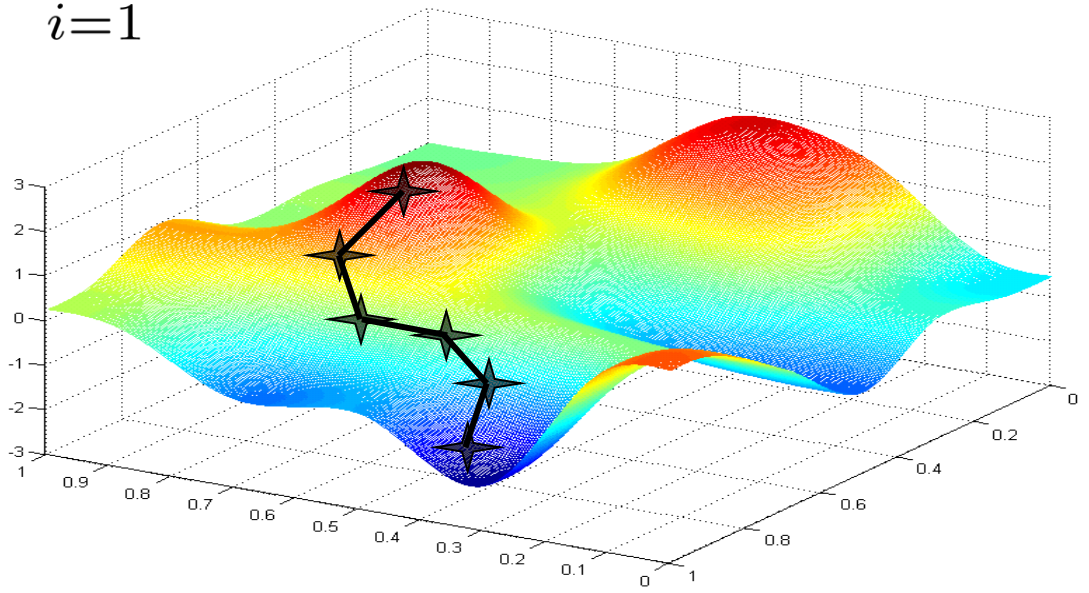
AI Background

- Training a model is:

Hypothesis: $h_\theta(x) = \theta_0 + \theta_1 x$

Cost Function: $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$

Goal: $\underset{\theta_0, \theta_1}{\text{minimize}}\ J(\theta_0, \theta_1)$

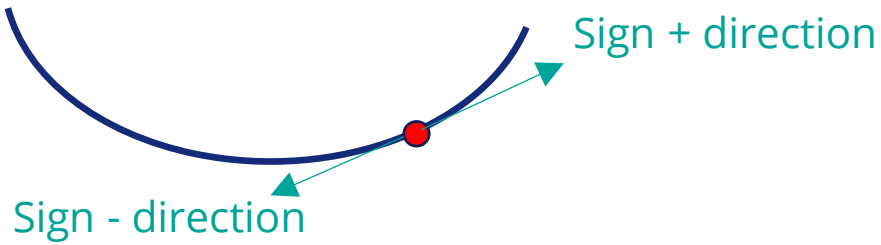


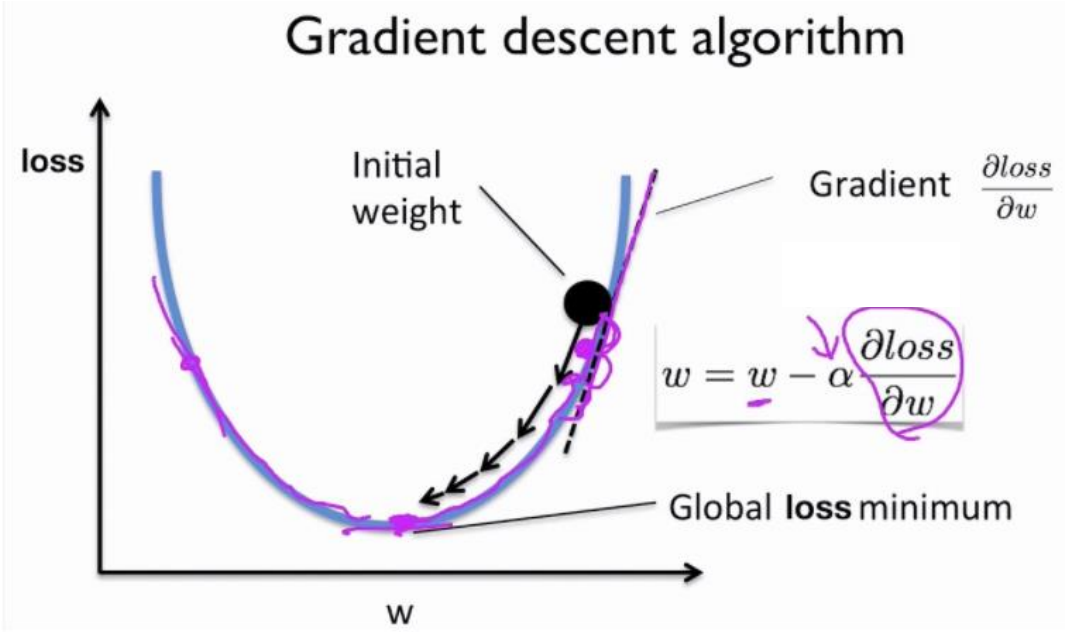
Copyright © 2025

Solving the training problem: GD

Cost Function: $$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

Goal: $$\underset{\theta_0, \theta_1}{\text{minimize}}\ J(\theta_0, \theta_1)$$





Copyright © 2025

Adversarial Attacks

Copyright © 2025

Adversarial Attacks

Carefully crafted imperceptible noise, that forces the model to misclassify when added to an input

$\boldsymbol{x}$

"panda"
57.7% confidence

$+ .007 \times$

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

$=$

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

Copyright © 2025

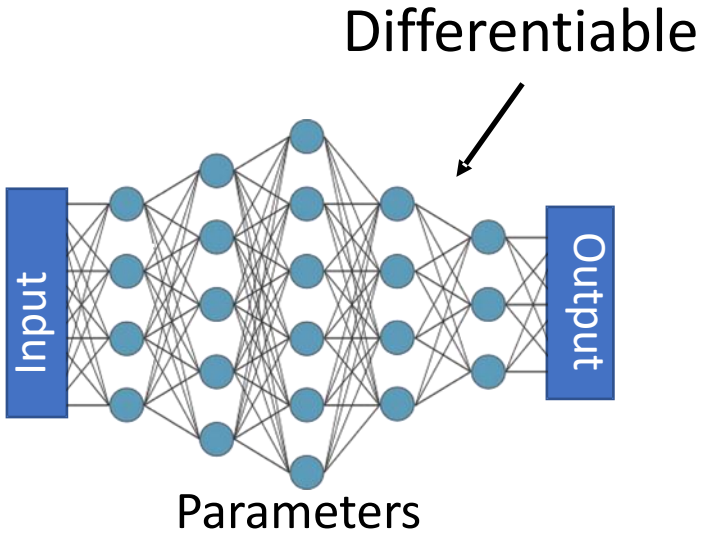
Where Do Adversarial Examples Come From?

To get an adv. example

~~Goal of training:~~ Model Parameters Input Correct Label

$$\min_{\theta} loss(\theta, x, y)$$

$$\max_{\delta} loss(\theta, x + \delta, y)$$



Can use gradient descent method to find good

Copyright © 2025

Attacker's goal (Untargeted)

- Given C, x, find x' such that C(x') ± C(x)
- x' should be semantically same as x
  - Approximation should be small. $\|x - x'\|_p$

- The cost should be acceptable
  - Number of queries should be small
  - Computational cost should be limited

Copyright © 2025

Attacker's goal (Targeted)

- Given C, x, and t, find x' such that C(x')=t
- X' should be semantically same as x
  - Approximation should be small. $\|x - x'\|_p$

- The cost should be acceptable
  - Number of queries should be small
  - Computational cost should be limited

Copyright © 2025

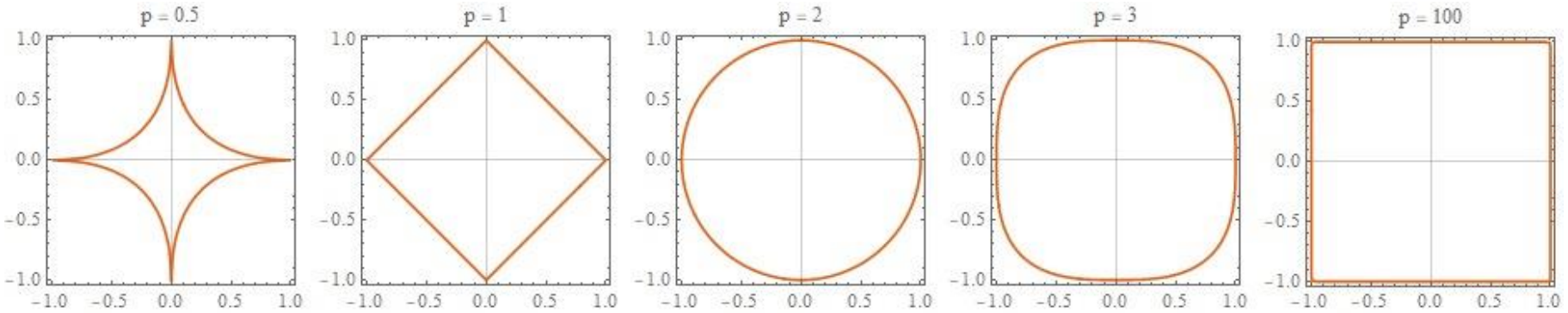
Measuring the noise budget

Typically, (convex) ℓ_p norms are used as constraints/regularizers

ℓ_p norms with $p \geq 1$, $\ell_p(\boldsymbol{x}) = \left(\sum_j |x_j|^p\right)^{1/p}$

Most popular examples

- ℓ_0 is not convex, and amounts to counting non-zero elements in x
- $\ell_1 = |x_1| + |x_2| + \cdots + |x_d|$
- $\ell_2 = x_1^2 + x_2^2 + \ldots + x_d^2$
- $\ell_\infty = \max_j |x_j|$



Copyright © 2025

Types of Adversarial Attacks

- Fast Gradient Sign Method (FGSM)
- Basic Iterative Method (BIM)
- Moment Iterative Method (MIM)
- Jacobian Saliency Map (JSMA)
- Deep Fool (DF)
- Carlini & Wagner (C&W)

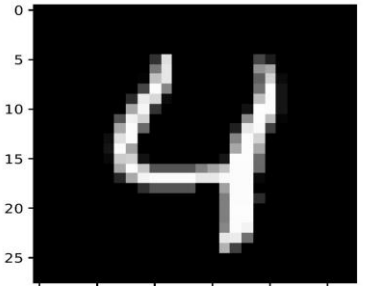
Copyright © 2025

FGSM

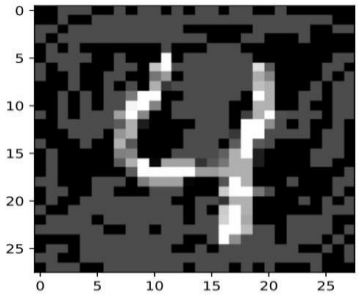
- Perturbs the image with gradient of the loss with respect to the input
- The FGSM adversarial perturbation generated is:

$$X^{adv} = X + \epsilon * sign(\nabla x\ loss(\theta, X, y_{true}))$$

- **Advantage:** Fast Attack, Single Iteration
- **Disadvantage:** Large Perturbation area required



Normal sample: 4



Classified as 6

Copyright © 2025

Real life?

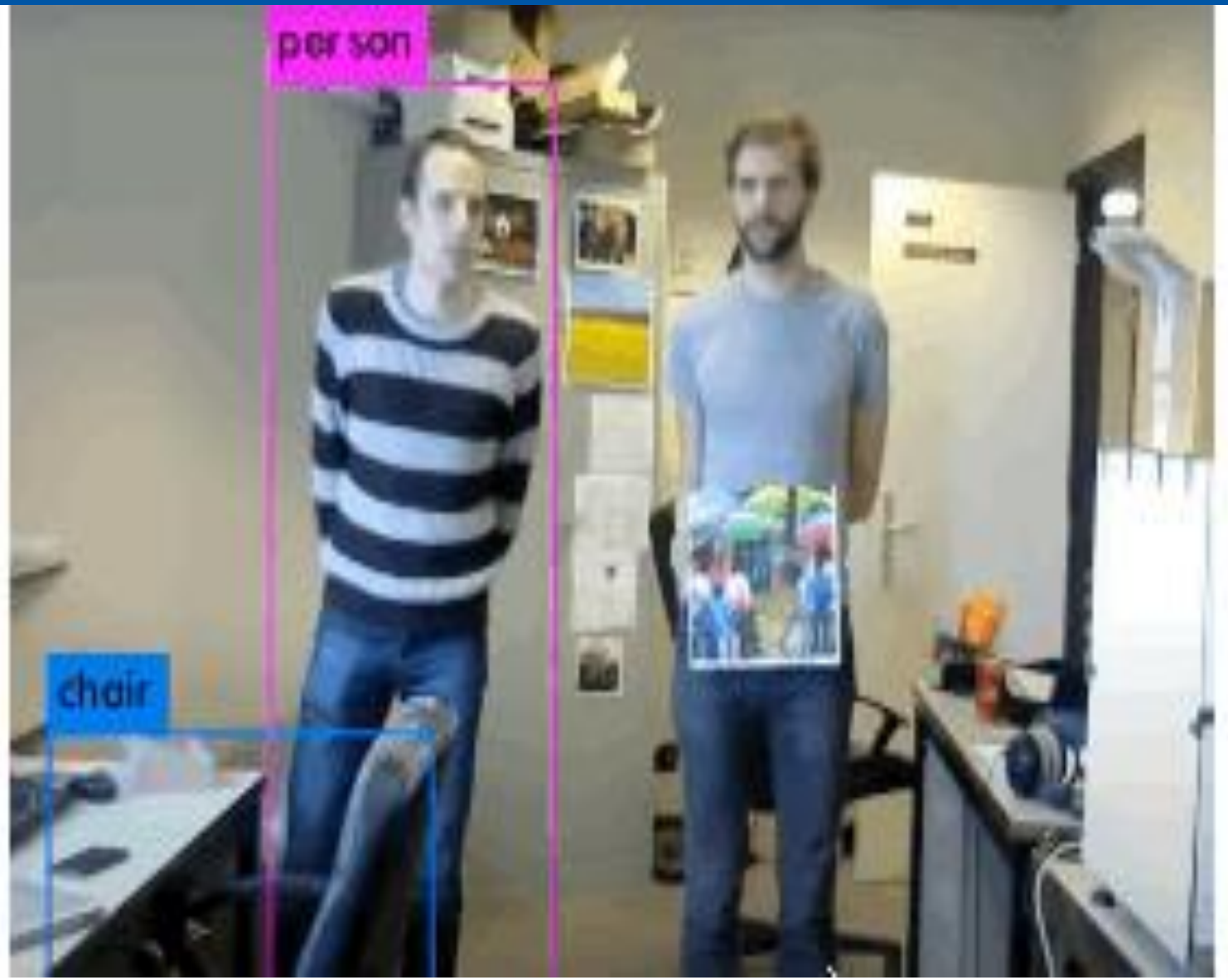
- Who cares about confusing pandas/digits?
- How to add noise in real life scenarios?

Copyright © 2025

Surveillance cameras?

- Space-constrained
- printable

⇒ Practicle in real-life



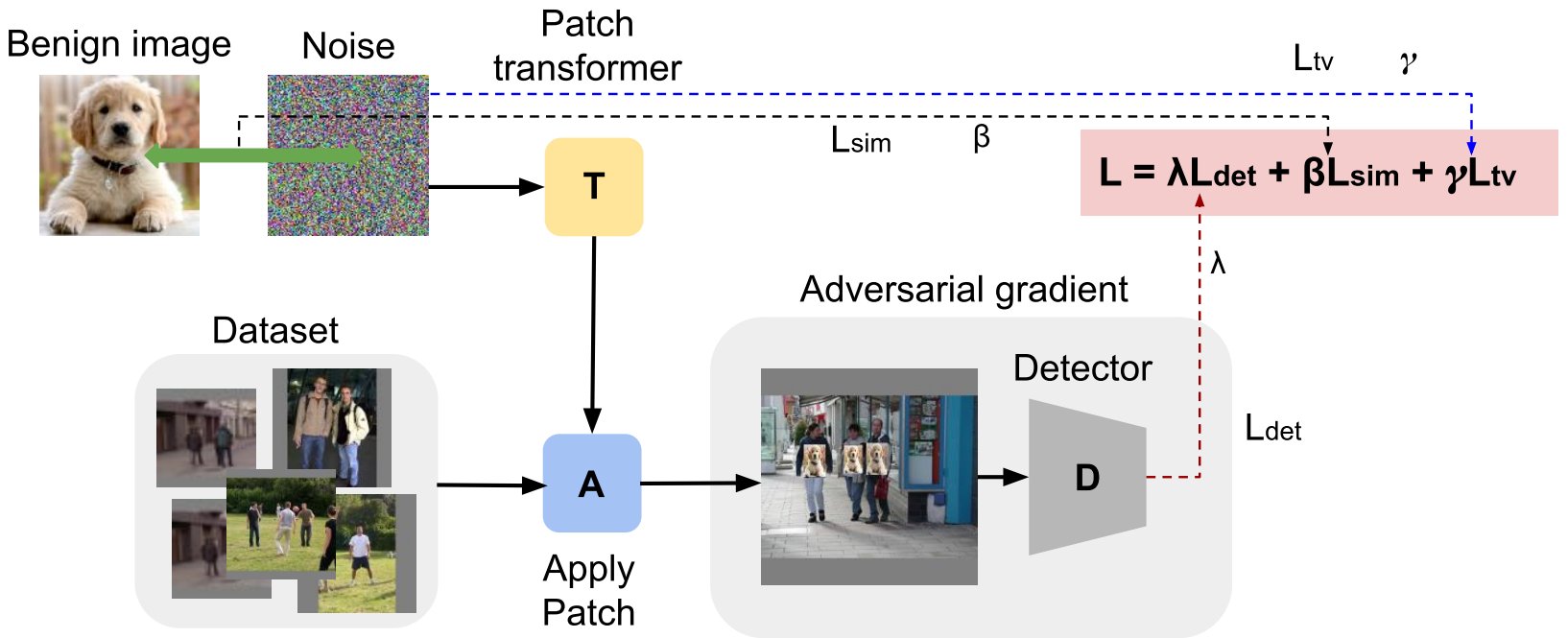
Copyright © 2025

Naturalistic Patch

Objective:
Design an adversarial patch that looks like a natural (benign) image.

Proposed method:

- Start with a noise and a target benign image
- Maximize the loss of the network, as well as the similarity between the noise and the target.



Copyright © 2025

Naturalistic Patch

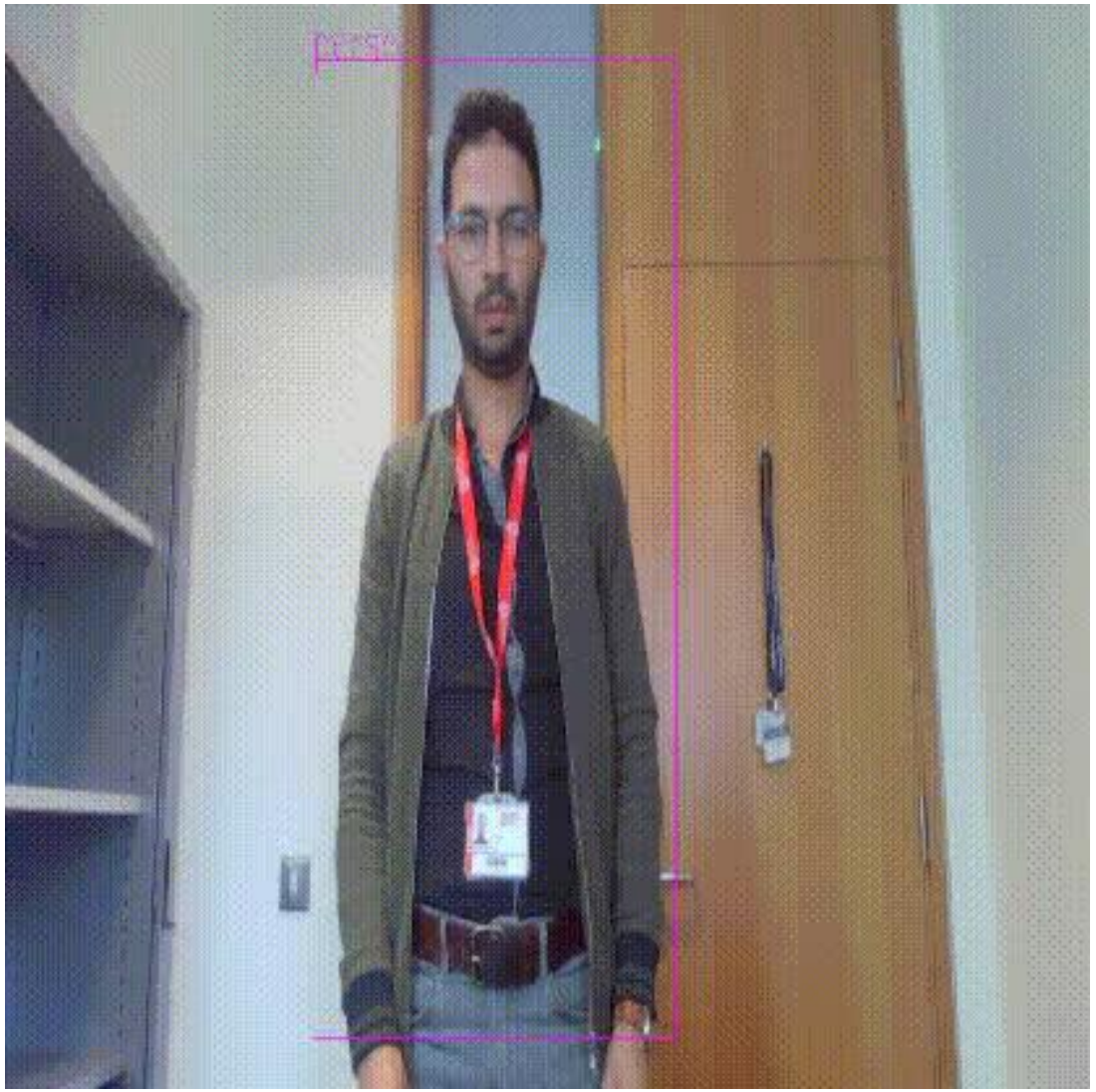
Experimental results: mean average precision (mAP)

Trained on \ Victim	Yolov2	Yolov3 tiny	Yolov3	Yolov4 tiny	Yolov4
(P1) yolov2	**29.80%**	35.67%	37.61%	36.20%	65.43%
(P2) yolov3-tiny	-	**15.82%**	52.77%	29.37%	58.24%
(P3) yolov3	-	41.46%	**34.93%**	37.46%	56.29%
(P4) yolov4-tiny	-	27.18%	41.36%	**17.7%**	65.84%
(P5) yolov4	-	60.70%	57.40%	47.46%	**12.48%**

P1 P2 P3 P4 P5

Copyright © 2025

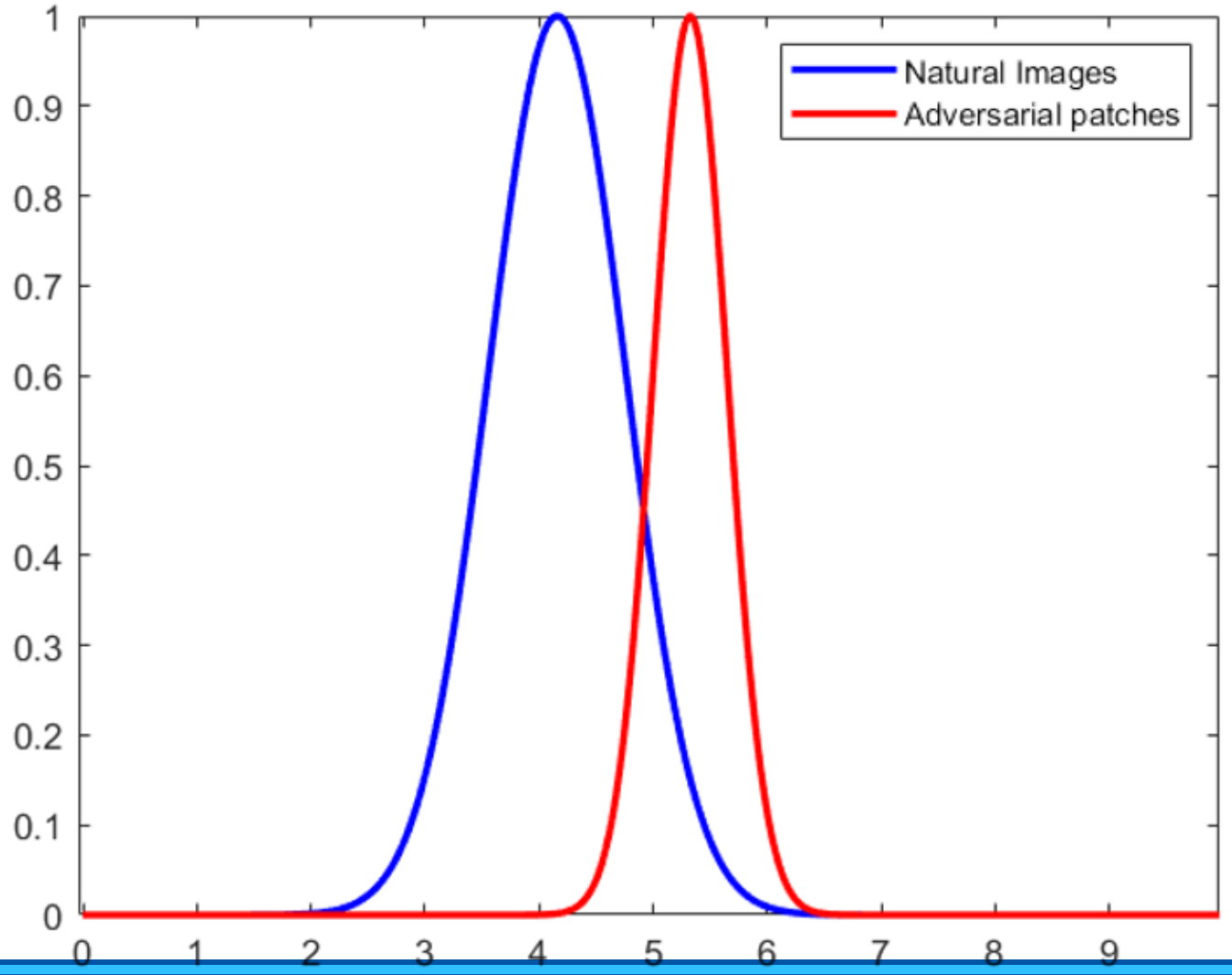
Naturalistic Patch



Copyright © 2025

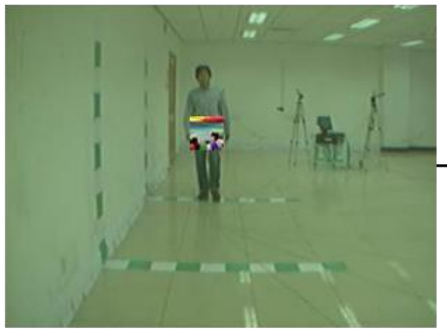
Jedi: Entropy-based defense

Observation:
Adversarial patches require a high concentration of information, hence entropy

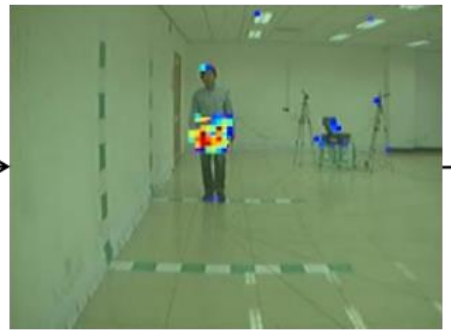


Copyright © 2025

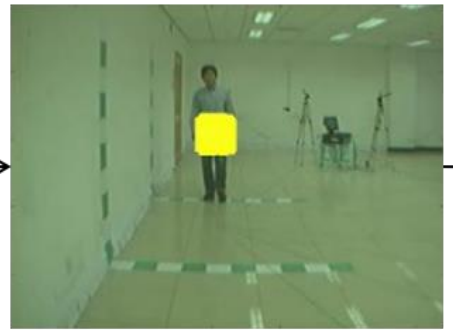
Jedi: Proposed Approach



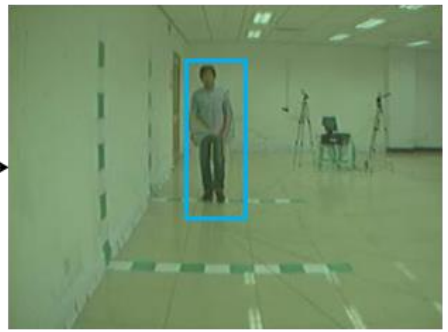
The person is undetected due to the adversarial patch



Step 1: We detect peaks in local entropy



Step 2: Using an autoencoder, we reconstruct an optimal mask that covers the adversarial patch while minimizing false positives



Step 3: The adversarial patch is removed by applying an inpainting method to the mask, and the original detection is restored

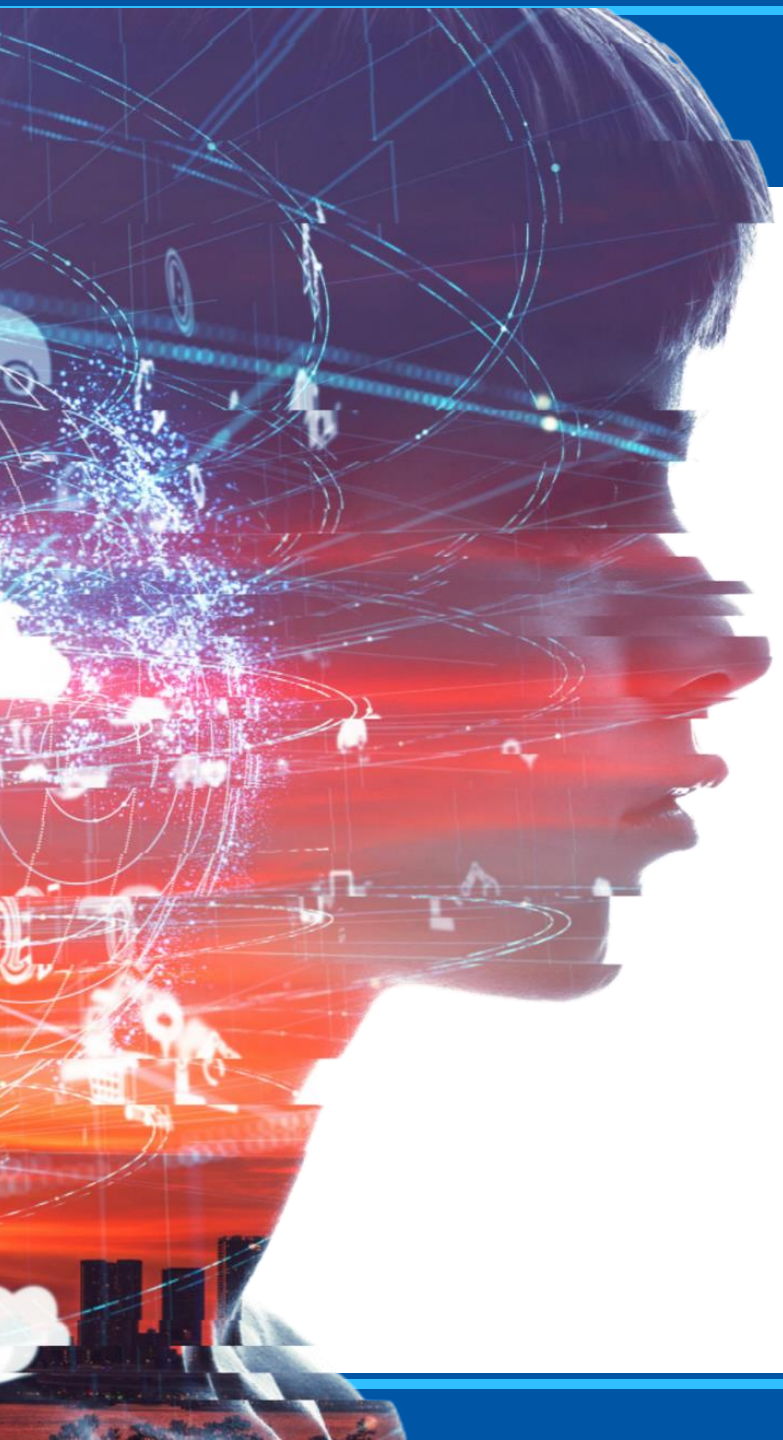
Copyright © 2025

Jedi: demo

Original attack

Jedi (Proposed defense)

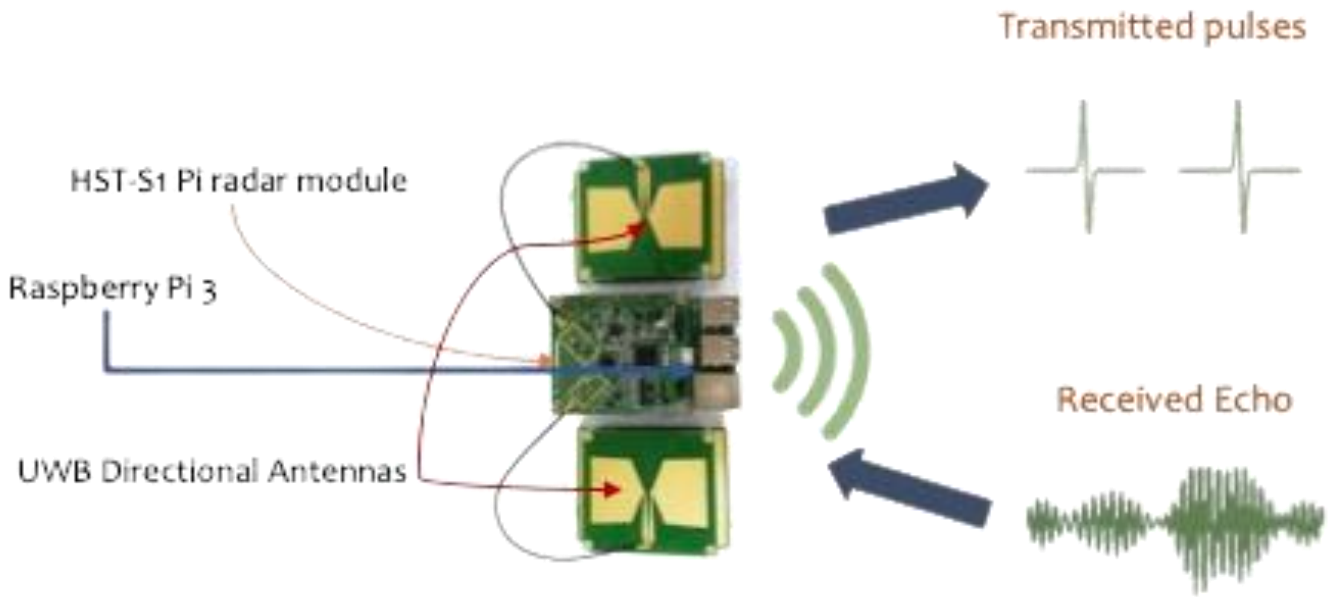
Copyright © 2025



Adversarial Attacks in the Wireless domain

Copyright © 2025

UWB Radar



The HST-D3 UWB Radar

- Embedded radar for short range vision

Copyright © 2025

UWB Radar: environment perception

System Design:

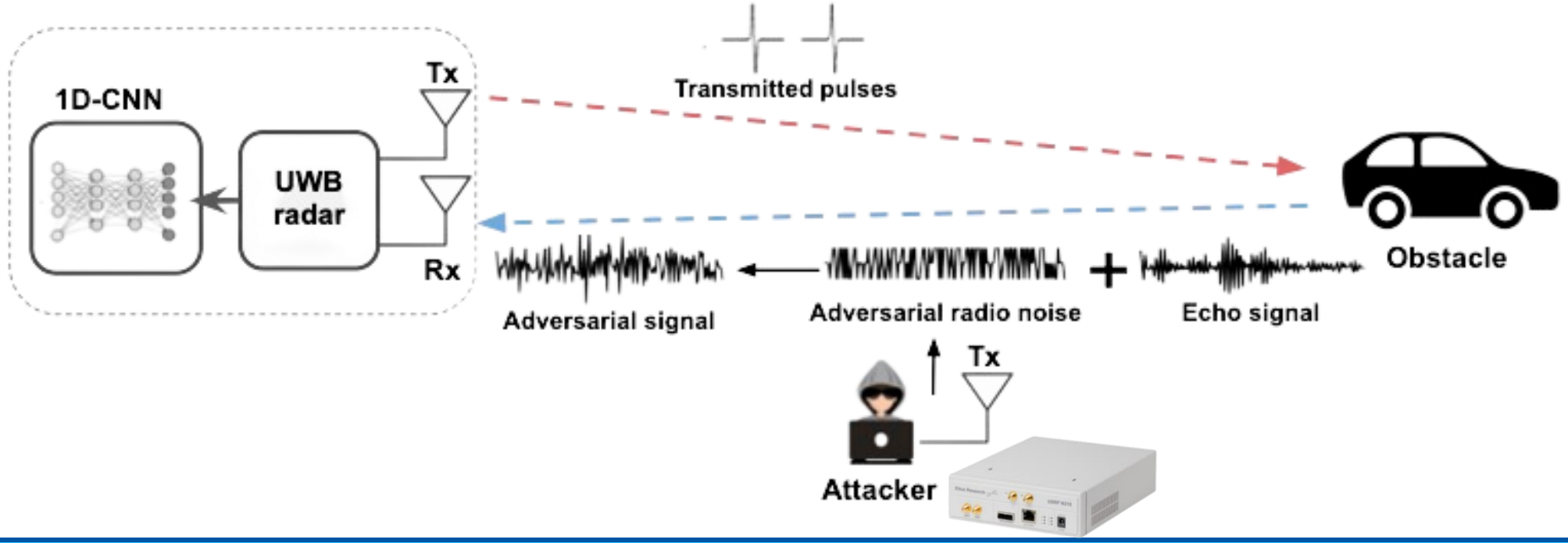
- UWB embedded radar for short range vision
- CNN for radar echo analysis, used to detect obstacles in the short range (especially in urban environments)
- High classification accuracy (97%)

Copyright © 2025

Threat Model

Assumptions:

- The adversary is physically in the surrounding of the UWB system
- The Adversary has access to a USRP device to inject arbitrary signals in the channel

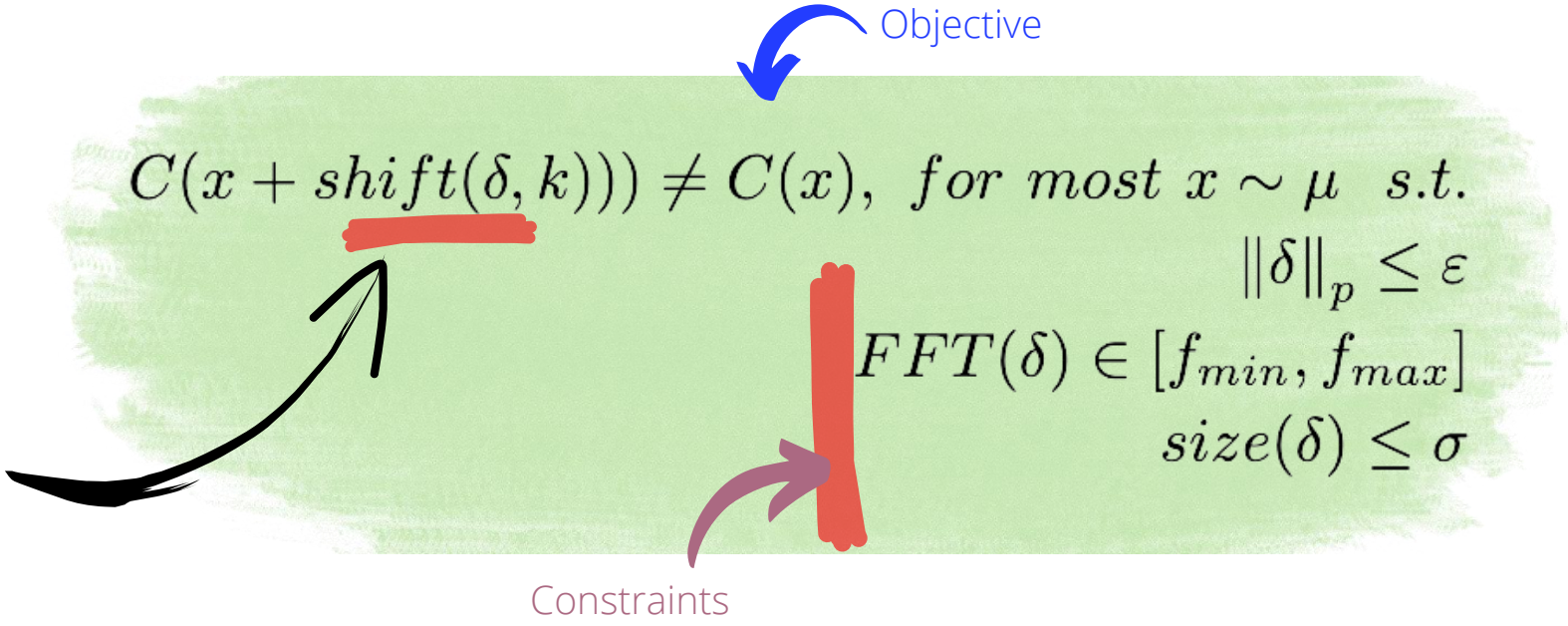


Copyright © 2025

Adversarial radio noise problem

We consider a defender that:

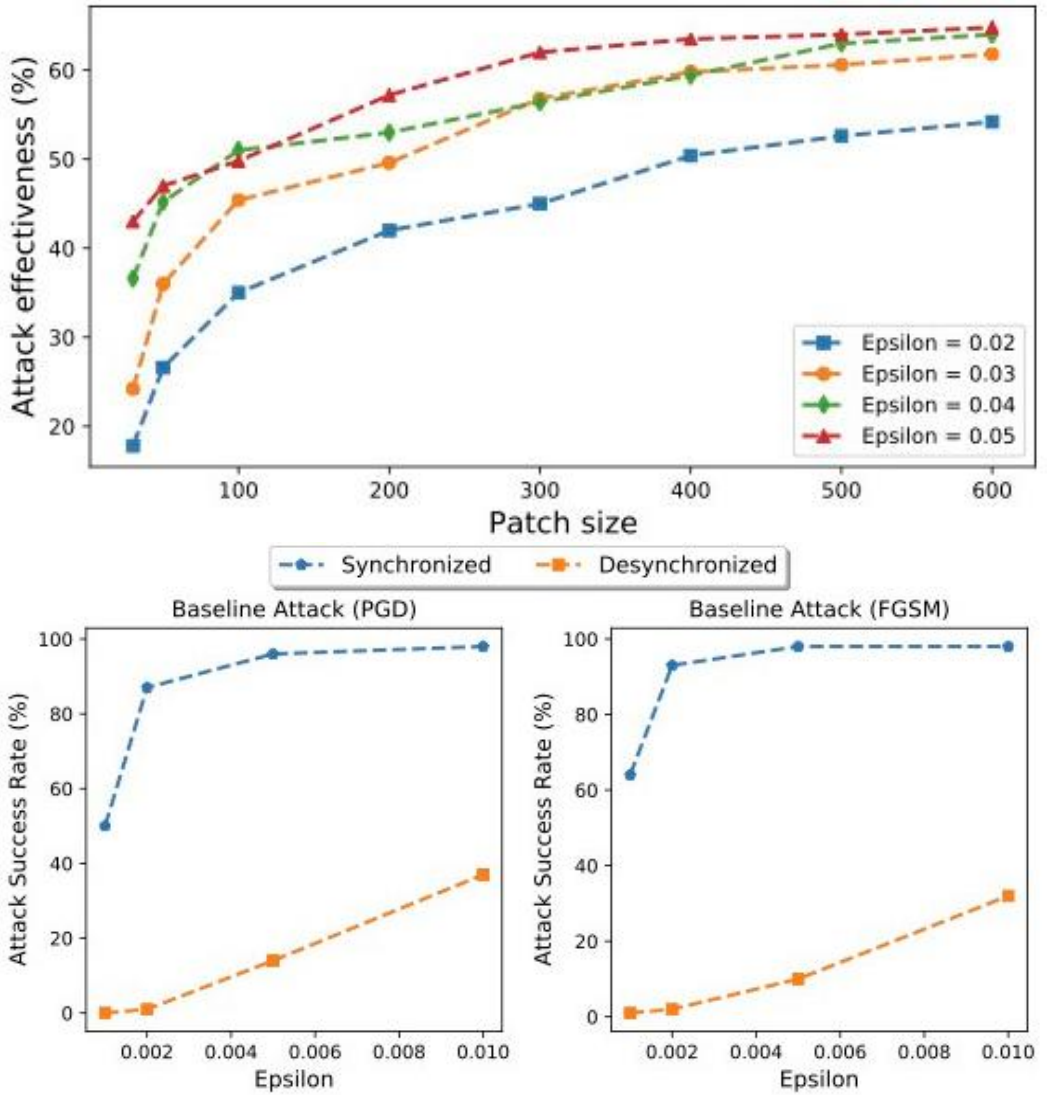
- Filters out unexpected frequencies
- Can potentially run spectrum sensing



Copyright © 2025

Adversarial Radar success rate

We test the attack under different patch magnitude/time budgets.
(600 is the frame size of an UWB Radar)

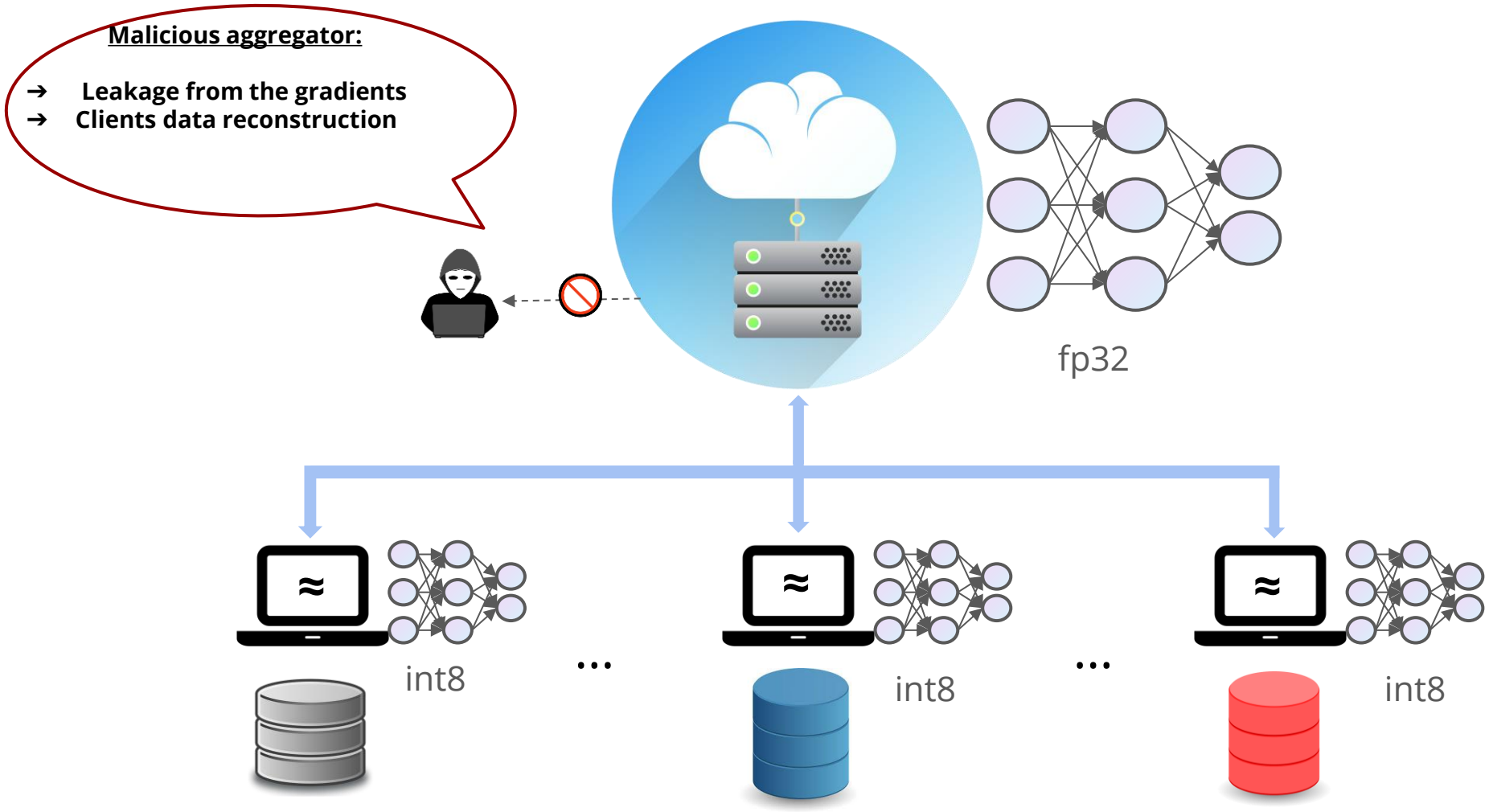


Copyright © 2025

From Security to Privacy

Copyright © 2025

Federated Learning Setting

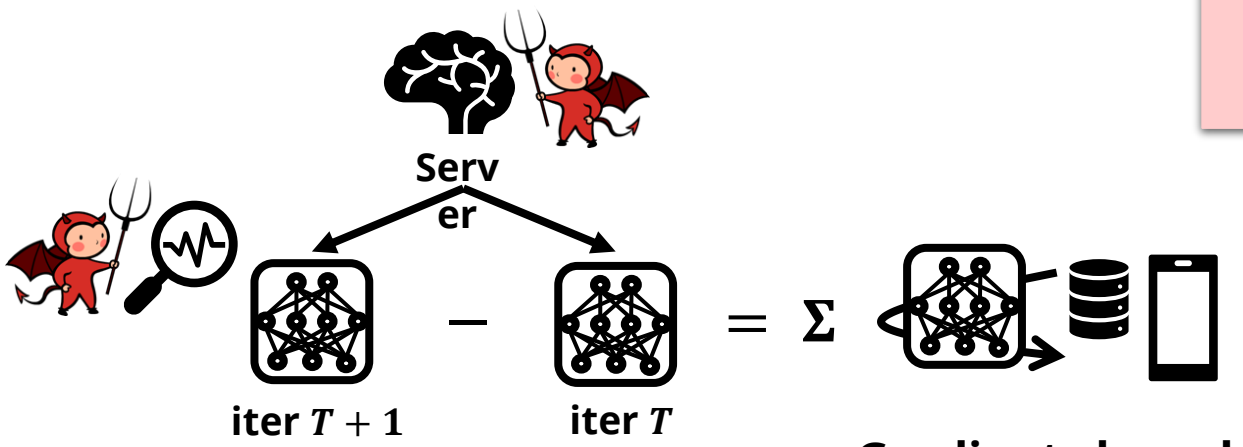


Copyright © 2025

Threat Model

Curious Server has access to model updates!

What information do updates leak?



Copyright © 2025

Attack: Deep Leakage from Gradients

- Exploits gradient updates shared between clients and a central server.

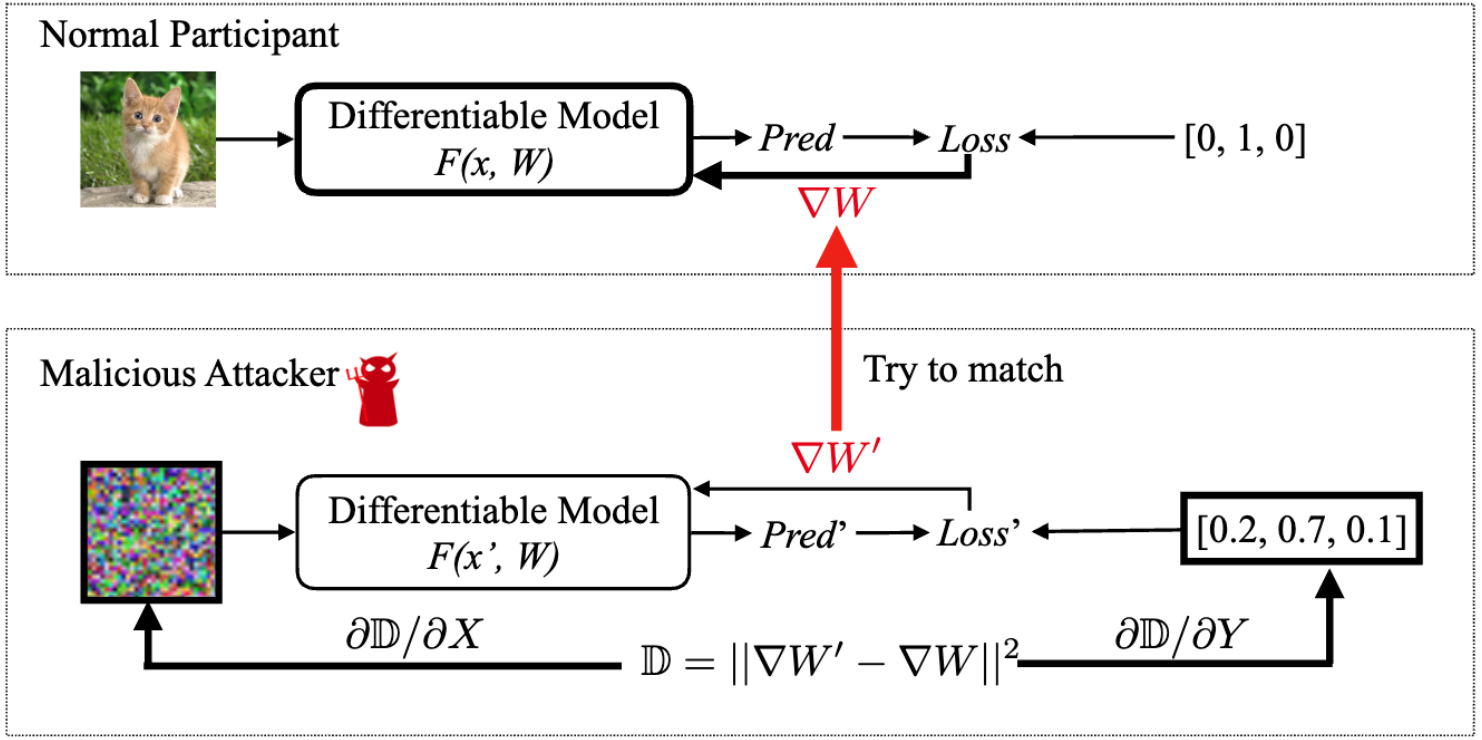
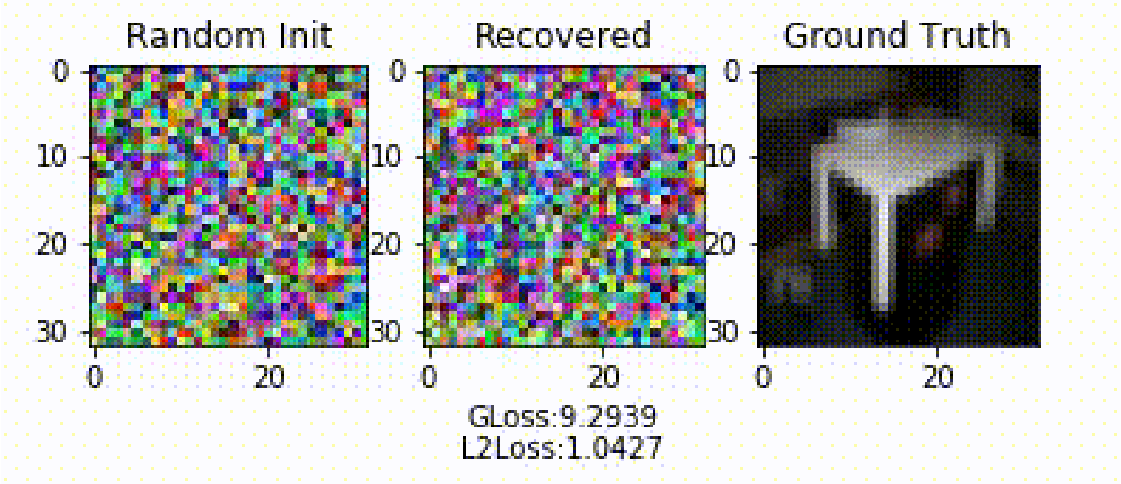
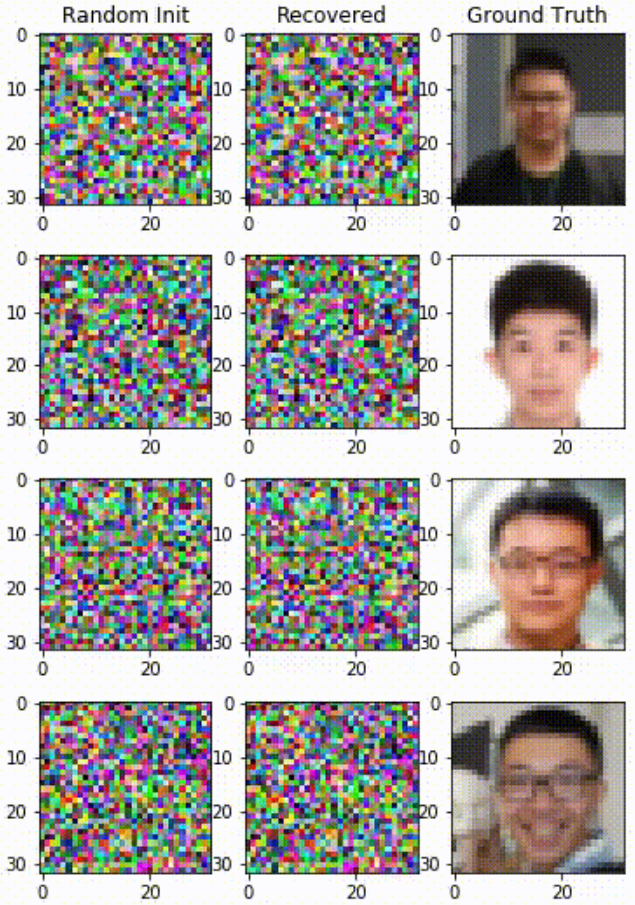


Figure 2: The overview of our DLG algorithm. Variables to be updated are marked with a bold border. While normal participants calculate ∇W to update parameter using its private training data, the malicious attacker updates its dummy inputs and labels to minimize the gradients distance. When the optimization finishes, the evil user is able to steal the training data from honest participants.

Copyright © 2025

DLG: Illustration of training data reconstruction





Copyright © 2025

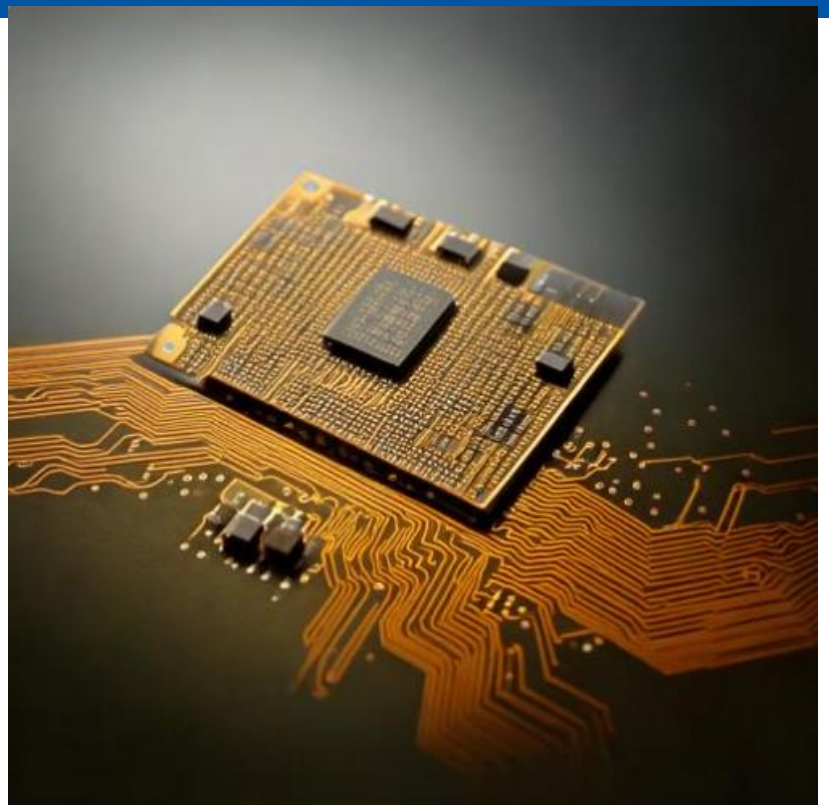
Edge-friendly AI privacy solutions

Copyright © 2025

Approximate computing

Approximate circuits are designed to trade off some degree of accuracy for gains in **power efficiency**, **speed**, and r**educed resource usage**.

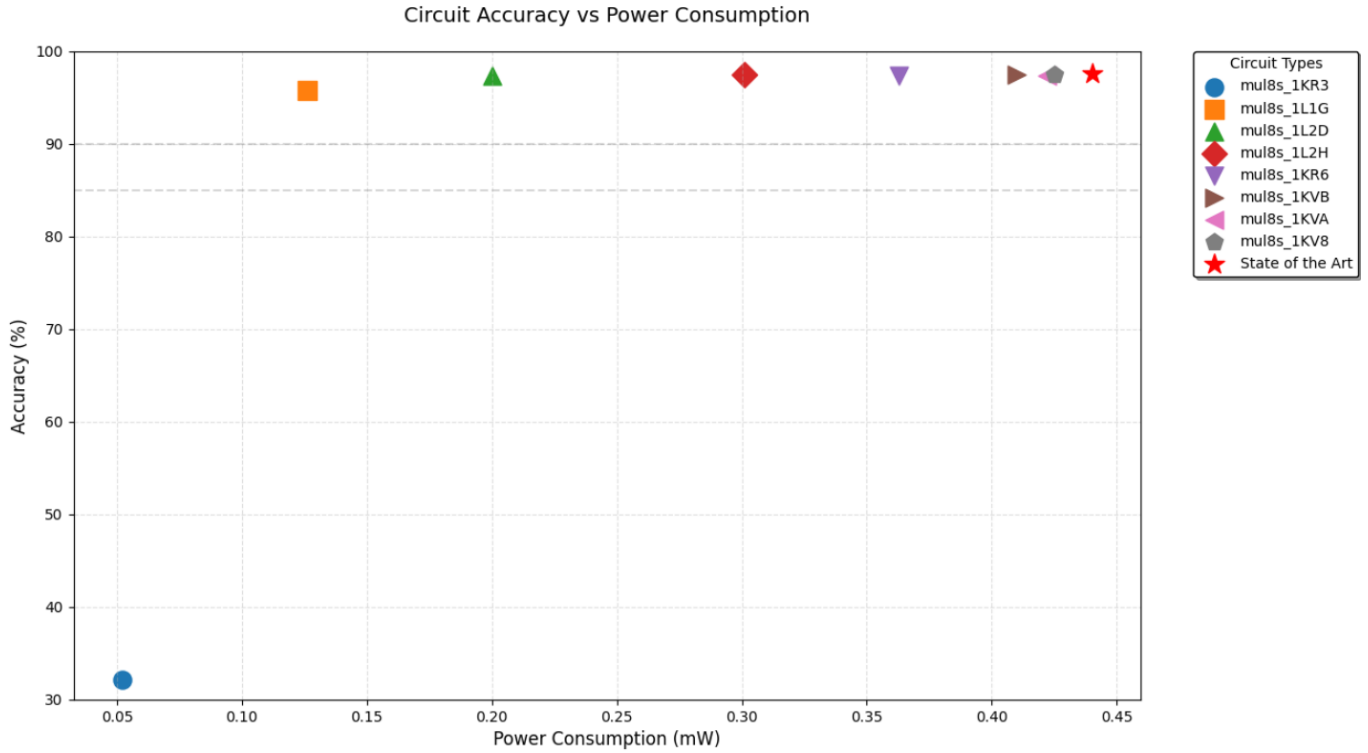
In federated learning, this concept is applied to optimize operations on resource-constrained edge devices.



Copyright © 2025

Energy/Accuracy Tradeoff

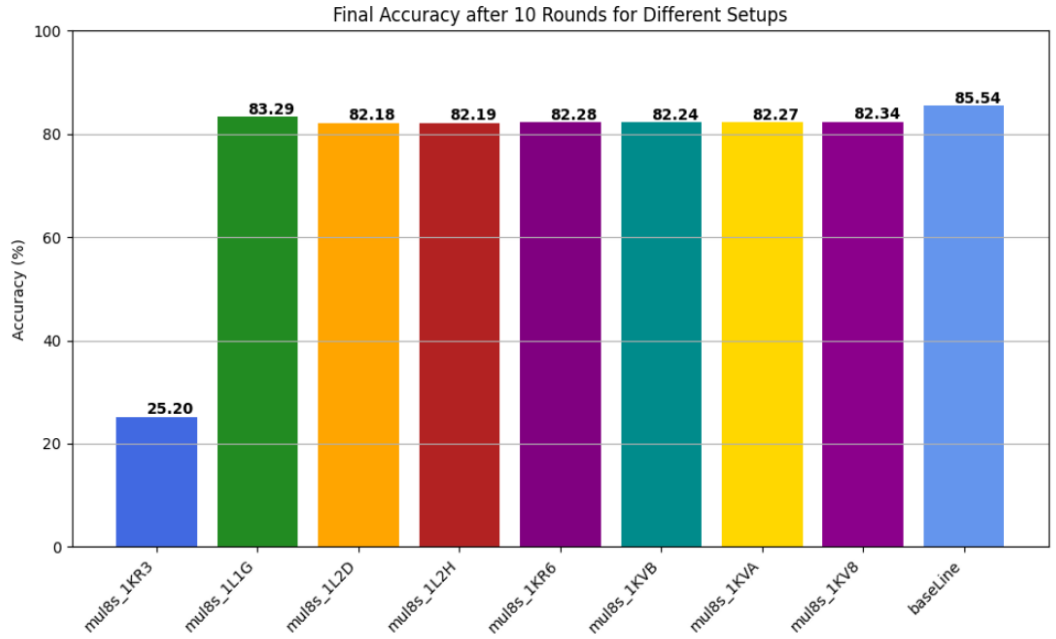
- MNIST accuracy variation under different Approximate multipliers (power consumption levels) under **conventional training setting**.



Copyright © 2025

QA-FL convergence

- State-of-the-art **convergence time**.
- State-of-the-art **Precision**
- By-product gain in both energy and communication bandwidth

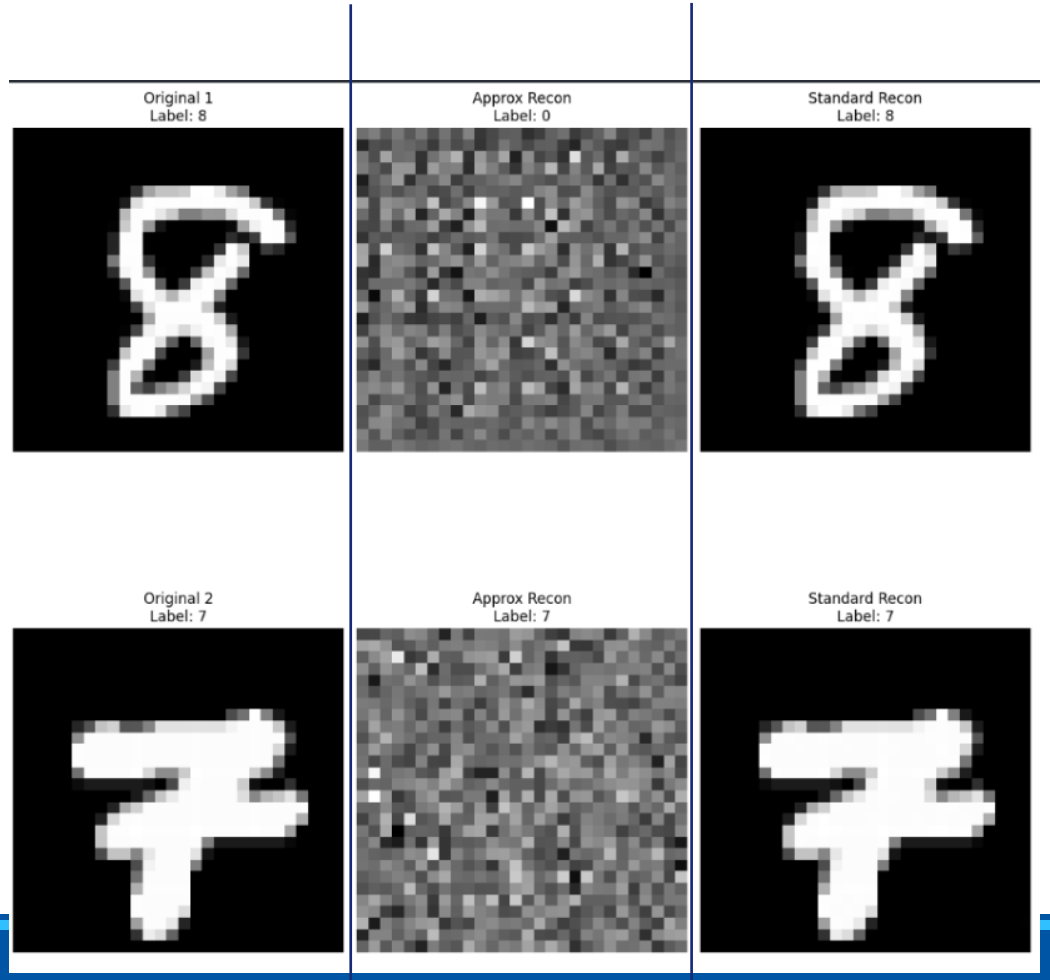


Copyright © 2025

Privacy Results (Qualitative)

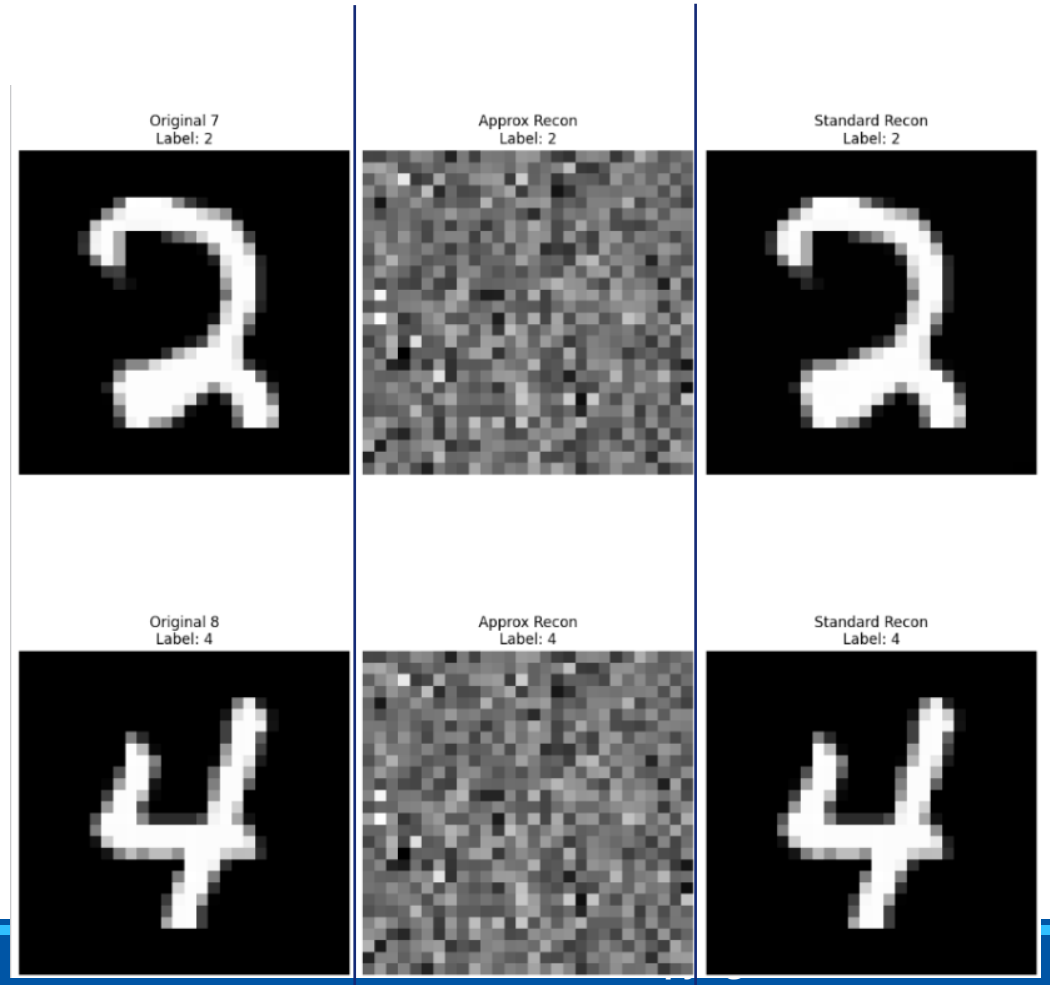
Original DLG(conventional)

DLG (AC)

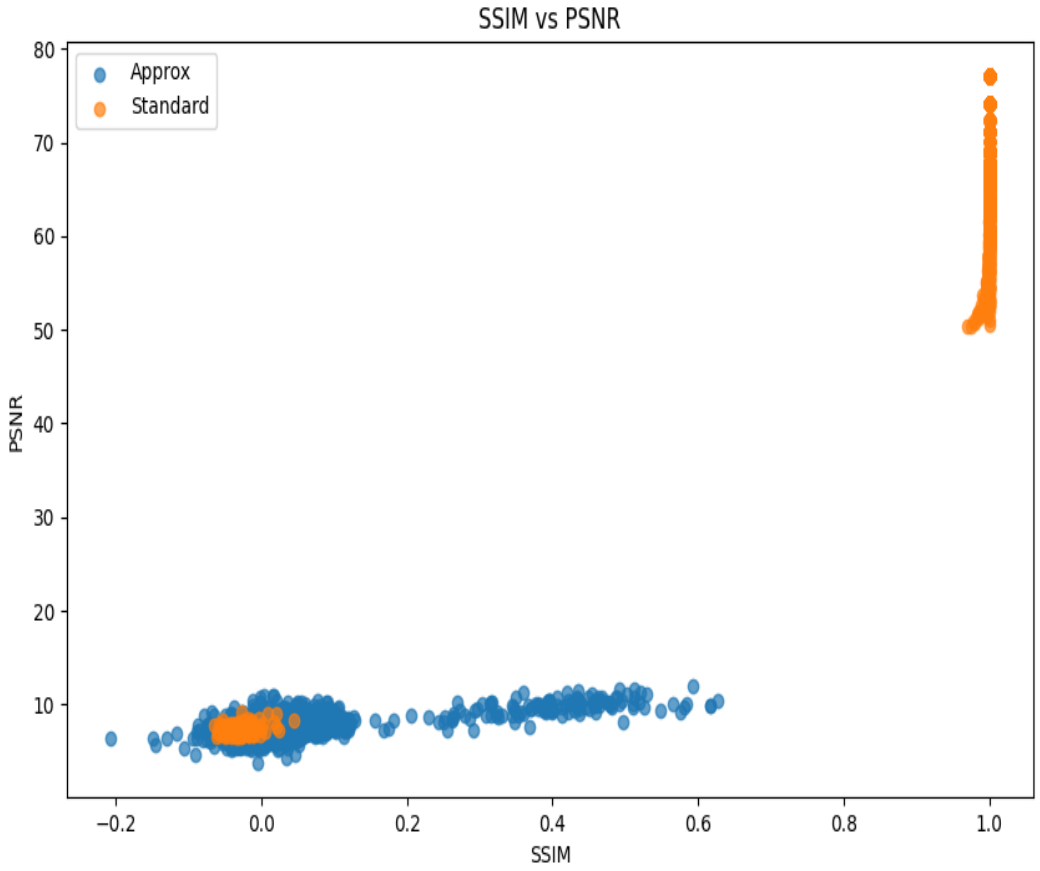


Original DLG(conventional)

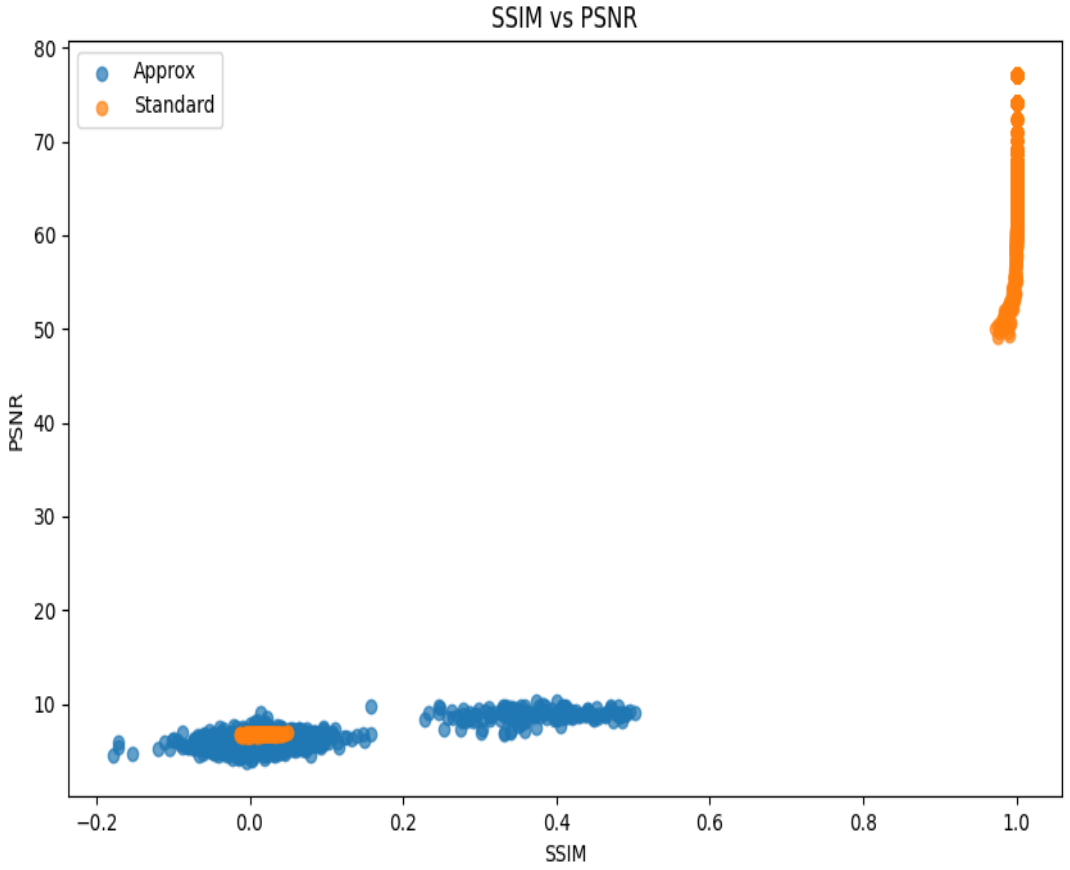
DLG (AC)



DLG Reconstruction accuracy (PSNR and SSIM)



Fashion MNIST

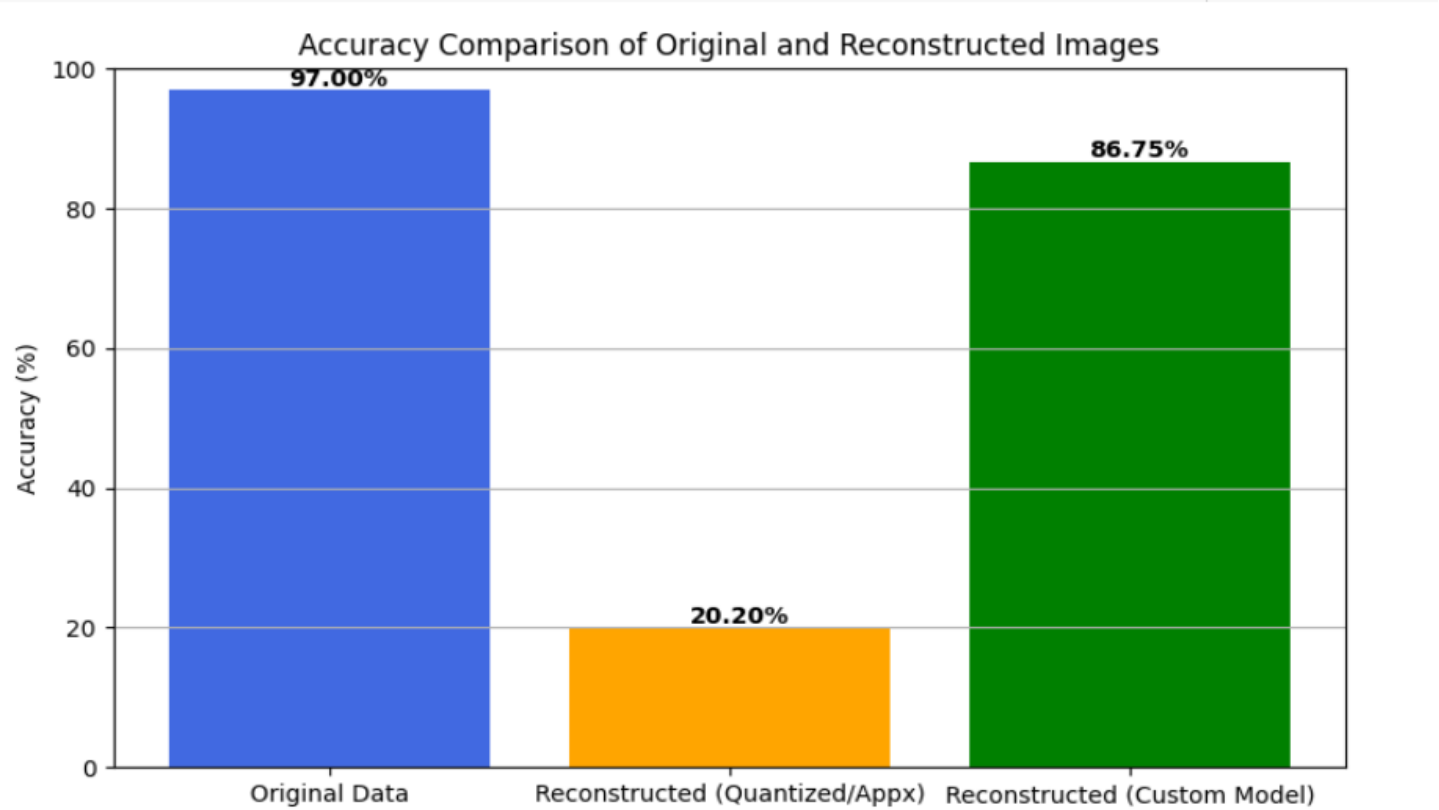


MNIST

Copyright © 2025

DLG Reconstruction accuracy

Can a pretrained Lenet-5 recognise the reconstructed images from the gradient?



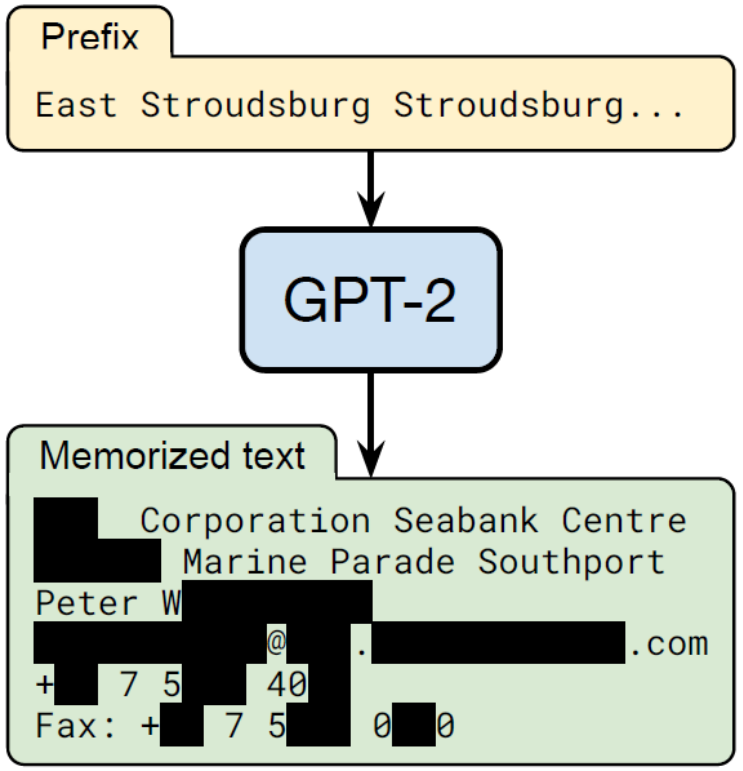
Copyright © 2025

Privacy Threat in Generative AI

Copyright © 2025

Verbatim memorization in LLMs…

- Example of training data extraction
  - The model outputs a block of text, which included the full name, phone number, email address, and physical address of the person
  - GPT-2 regurgitated data memorized during the training process



Conclusion

- AI Security is NOT (just) a theoretical problem, and has high potential to cause considerable harm in real life and hence needs to be taken very seriously
- Edge AI Security has higher potential of defenses that are not available to the cloud-based AI

Copyright © 2025

Thank You

i.alouani@qub.ac.uk

EAI

https://edge-ai-tech.eu/

The two-day innovative educational initiative is designed for students, young researchers, and professionals eager to expand their understanding and skills in these cutting-edge technologies.

Participants will engage in interactive discussions, hands-on sessions, and lectures covering foundational concepts and innovative applications of edge AI across various domains with leading experts, gain insights into real-world applications, and work on team-based projects to foster collaboration and innovation.

Copyright © 2025