Tempest

ACADEMY

Conference

2023

Attacking and Defending Machine Learning-based Systems

Prof. Paulo Freitas
pfreitas@cin.ufpe.br

- Graduação
  - Engenharia Eletrônica @UFPE
  - University of Massachusetts
- Mestrado
  - Ciência da Computação @UFPE
- Doutorado
  - Computer Science @UFPE
  - Eng Elétrica @ÉTS
- Intrepid Control Systems
- Ericsson
- Microsoft Research
- Tempest Security Intelli

Tempest
ACADEMY
Conference

01 The AI Spring

02 The Risk

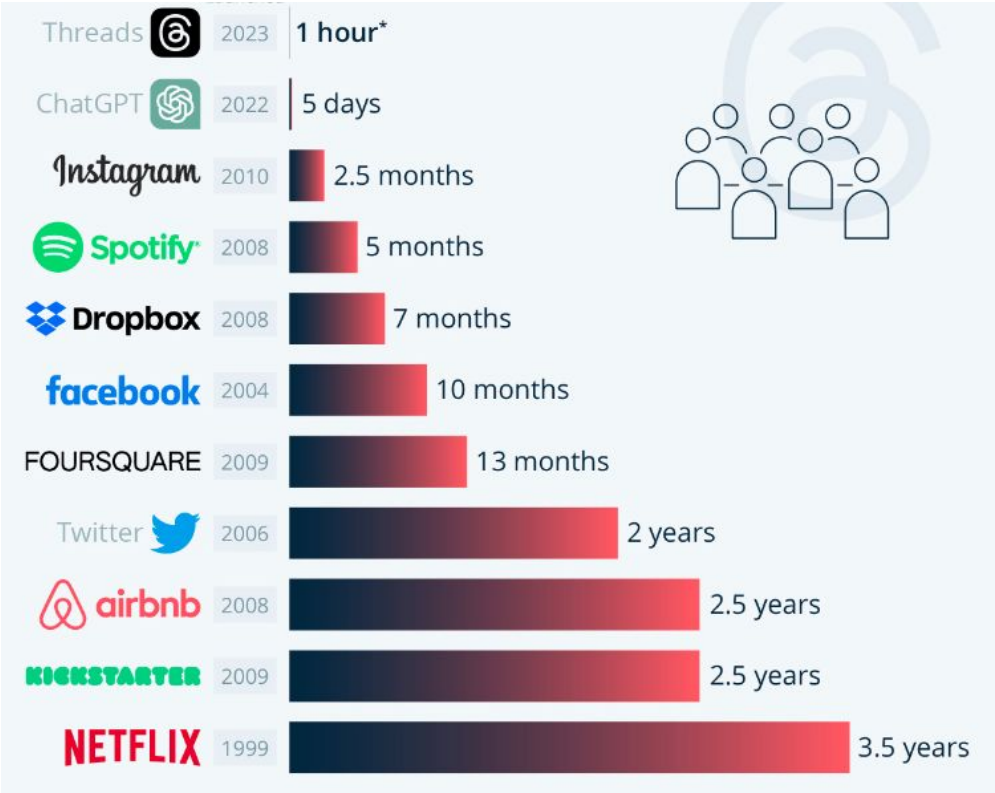
03 The Aftermath

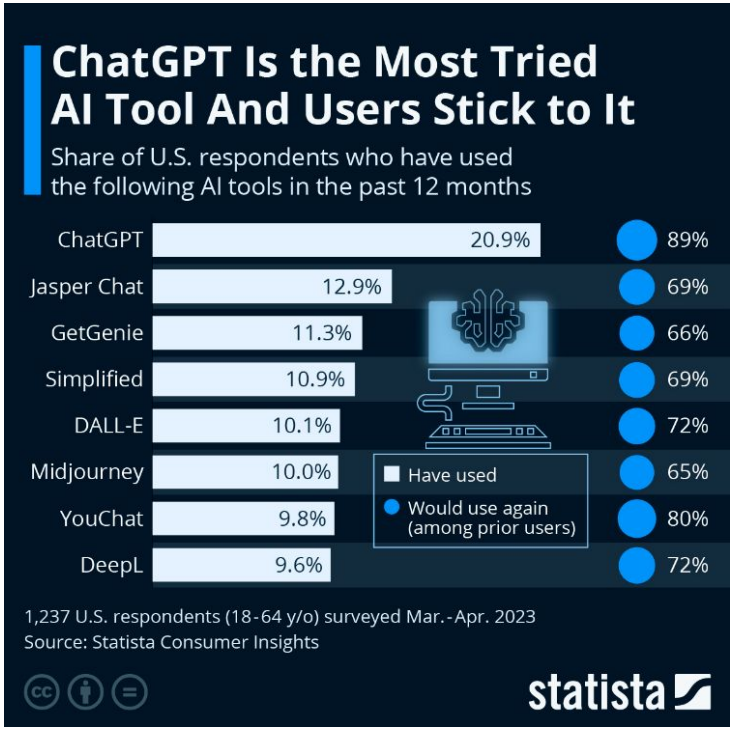
04 The Future

The AI Spring

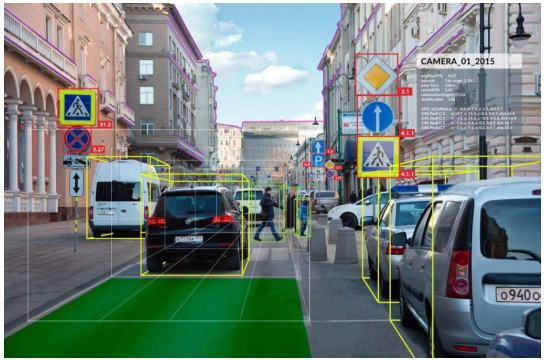
The AI Spring

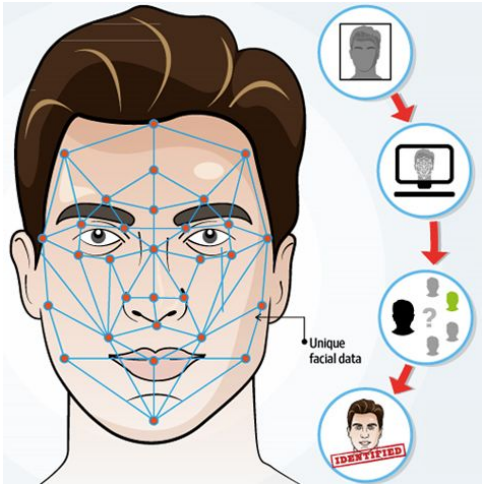
Tempest ACADEMY Conference





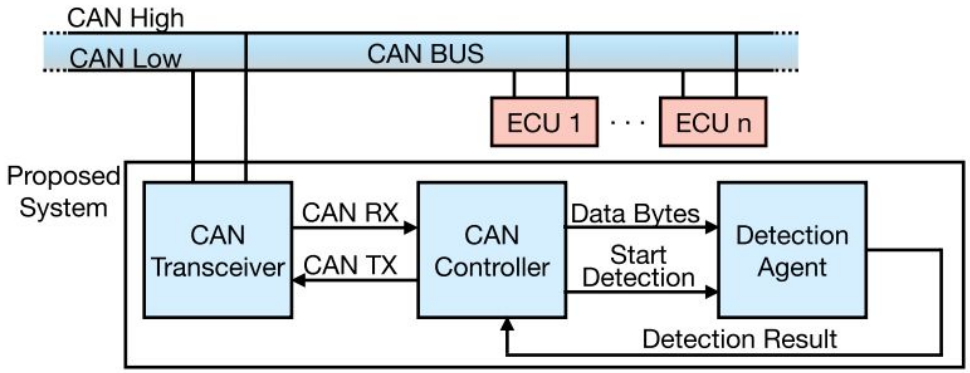
The AI Spring





The AI Spring

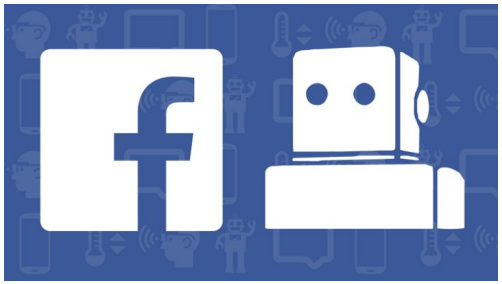
Tempest
ACADEMY
Conference



The AI Spring

Tempest
[ACADEMY]
Conference

Google
Machine Learning



100 STARTUPS USING ARTIFICIAL INTELLIGENCE TO TRANSFORM INDUSTRIES

CONVERSATIONAL AI/ BOTS
MindMeld, semantic machines, Maluuba, Mobvoi 出门问问, KITT.AI, snips, clara, AUTOMAT, X.

VISION
clarifai, Chronocam, Orbital Insight, pilot.ai, Captricity, GROKSTYLE

AUTO
nuTonomy, drive.ai, AIMOTIVE, nauto, nexar, ZOOX

ROBOTICS
UBTECH, ANKI, Rokid, dispatch

CYBERSECURITY
CYLANCE, sift science, sparkcognition, deepinstinct, Shift Technology, DARKTRACE

BUSINESS INTELLIGENCE & ANALYTICS
DataRobot, TRIFACTA, rapidminer, tamr, SIGOPT, Paxata, context relevant, Dataminr, CrowdFlower, logz.io

AD, SALES, CRM
TalkIQ, deep gram, [PERSADO], Qqpier, CHORUS, INSIDE SALES.COM, drawbridge, DigitalGenius, RESCI

CORE AI
affectiva, PETUUM, ALGORITHMIA, H_2O, loop.ai Labs COGNITIVE COMPUTING, CognitiveScale, sentient, Voyager Labs, Numenta, SI SCALED INFERENCE, skymind, vicarious, Digital Reasoning, bonsai, AYASDI

HEALTHCARE
freenome, CLOUD MEDX, zebra, enlitic, twoXAR, iCarbonX, Atomwise, deep genomics, BabyLon, BenevolentAI, Lunit

AI 100

THE INNOVATION SUMMIT
CBINSIGHTS

TEXT ANALYSIS/ GENERATION
textio, fido.ai, cortical.io, NarrativeScience

IOT/IIOT
nanit, KONUX, VERDIGRIS, SIGHT MACHINE Manufacturing Analytics

COMMERCE
bloomreach, mode.ai

FINTECH & INSURANCE
CAPE ANALYTICS, KENSHO, NUMERAI, alphasense, Kasisto

OTHER
GIGSTER, prospera, Citrine, BLUERIVER TECHNOLOGY, ROSS, zymergen, Descartes Labs, gradescope, talla

Tempest
[ACADEMY]
Conference

But why do we use AI at all?

Why not stick to the “classic” algorithms?

Tempest
ACADEMY
Conference

Let’s dive into an example...

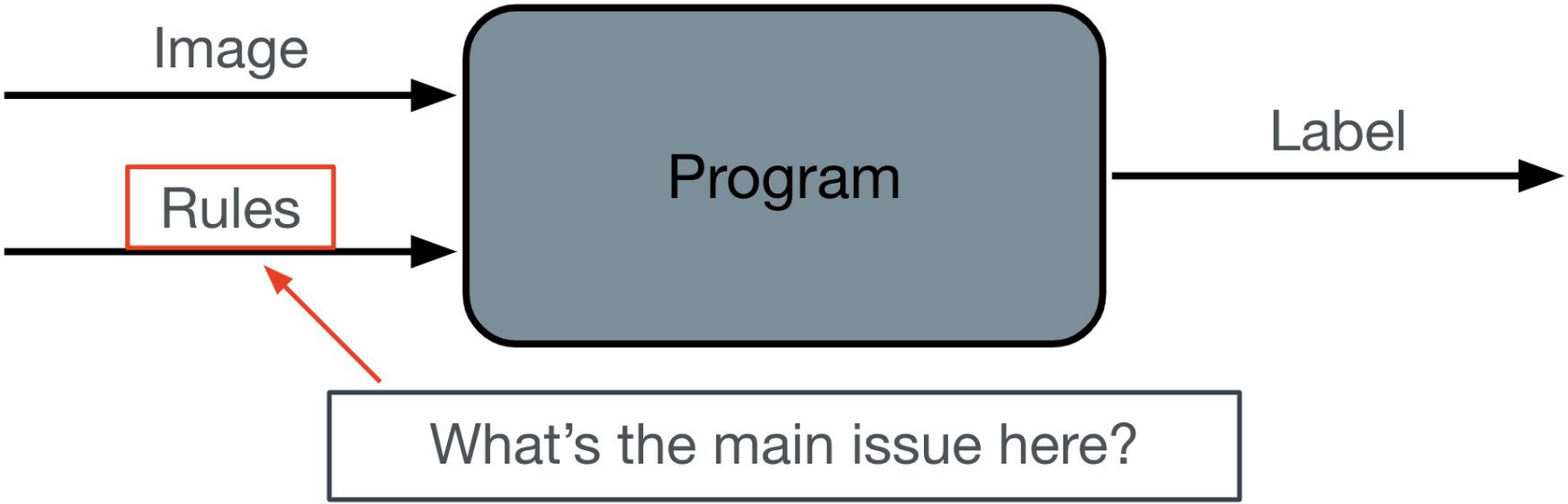
Let’s assume we have the following task: given an arbitrary image, decide if it is a panda or not.

Is this a Panda?

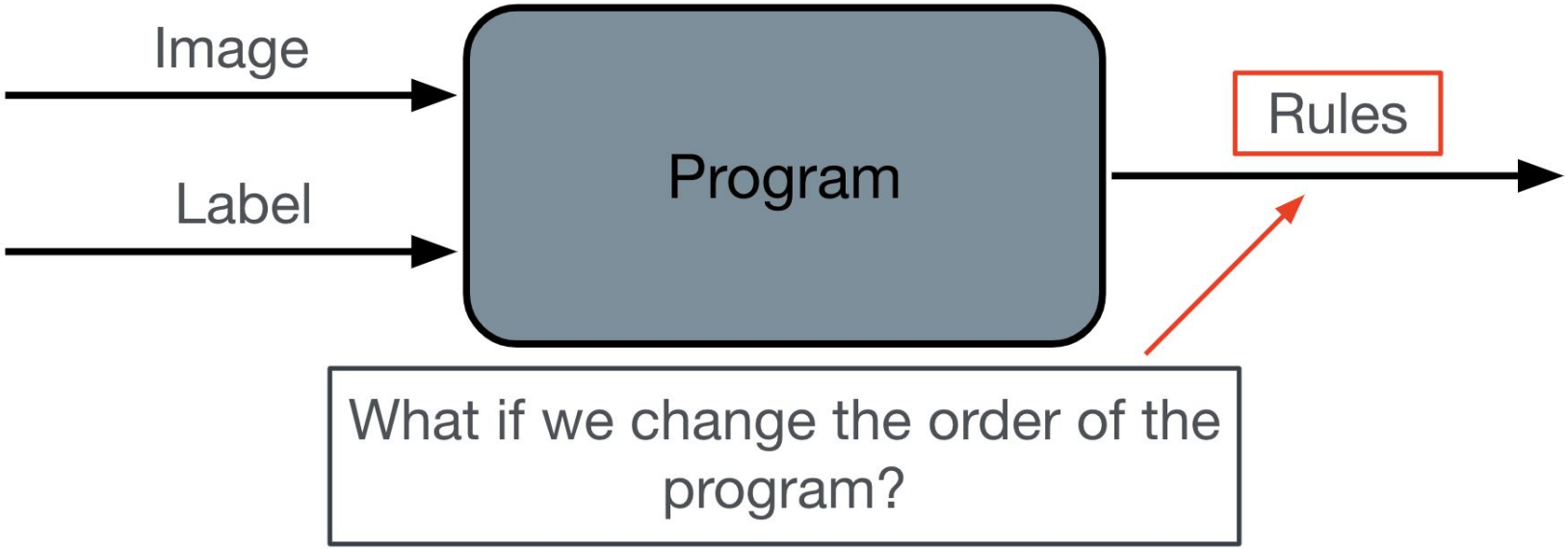
Tempest
ACADEMY
Conference

- Read each pixel;
- Analyze the pixels for the colors;
- Look for round black shapes;
- Find black heart-shaped figure just bellow and between the eyes;
- Define a texture that matches a panda skin;
- Remove other objects;
- Etc.

Tempest
ACADEMY
Conference

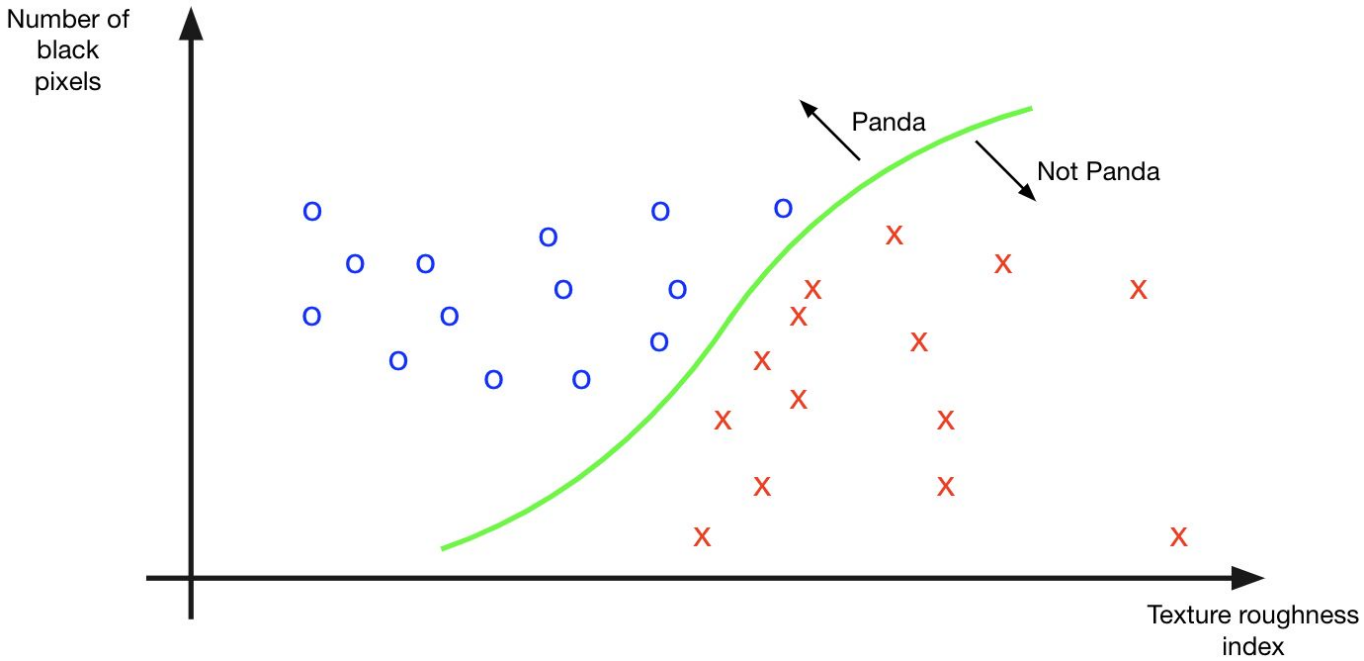


Tempest
ACADEMY
Conference



Fronteira de Decisão

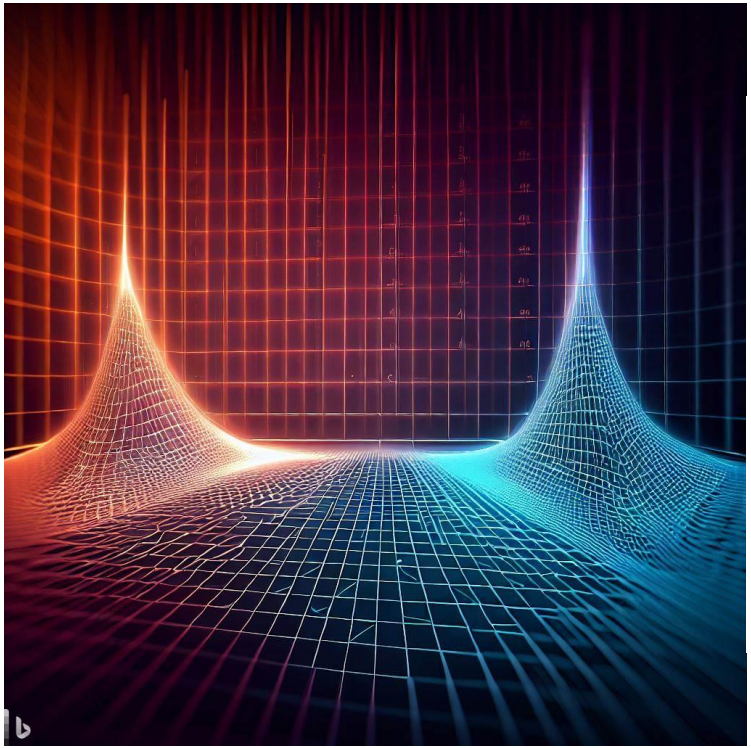
Tempest
[ACADEMY]
Conference

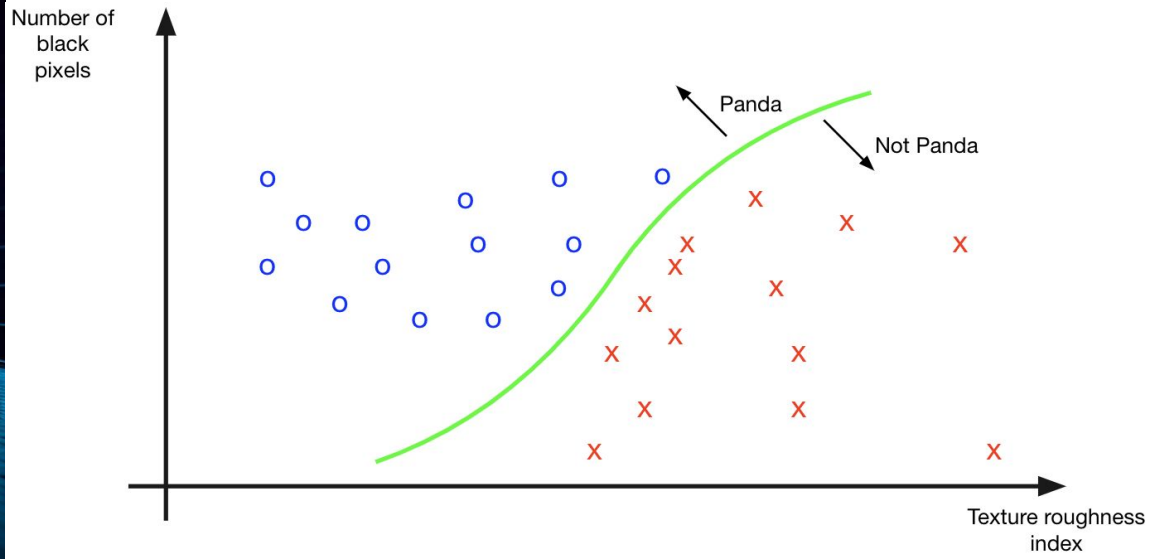


Fronteiras de decisão

Fronteira de Decisão

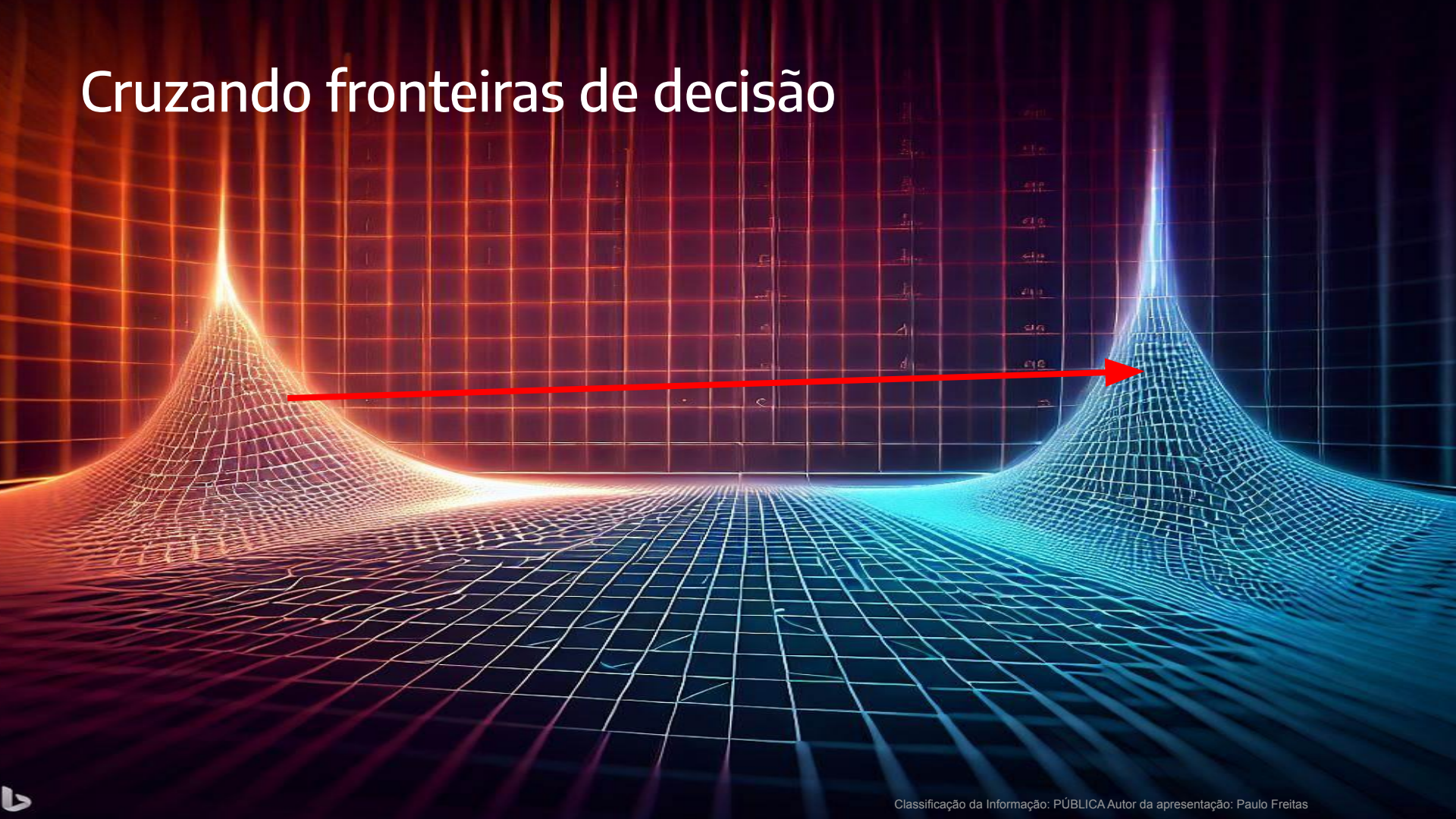
Tempest
[ACADEMY]
Conference





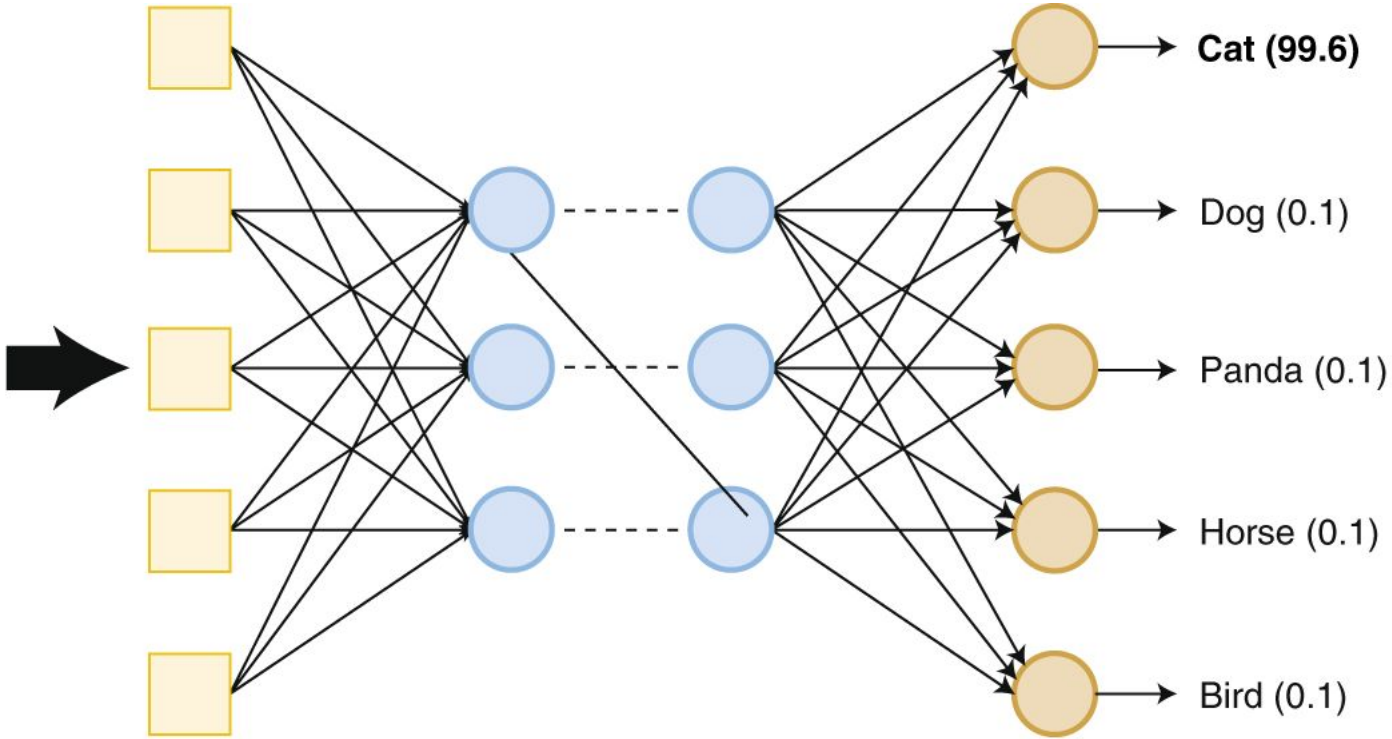
The Risk

Cruzando fronteiras de decisão



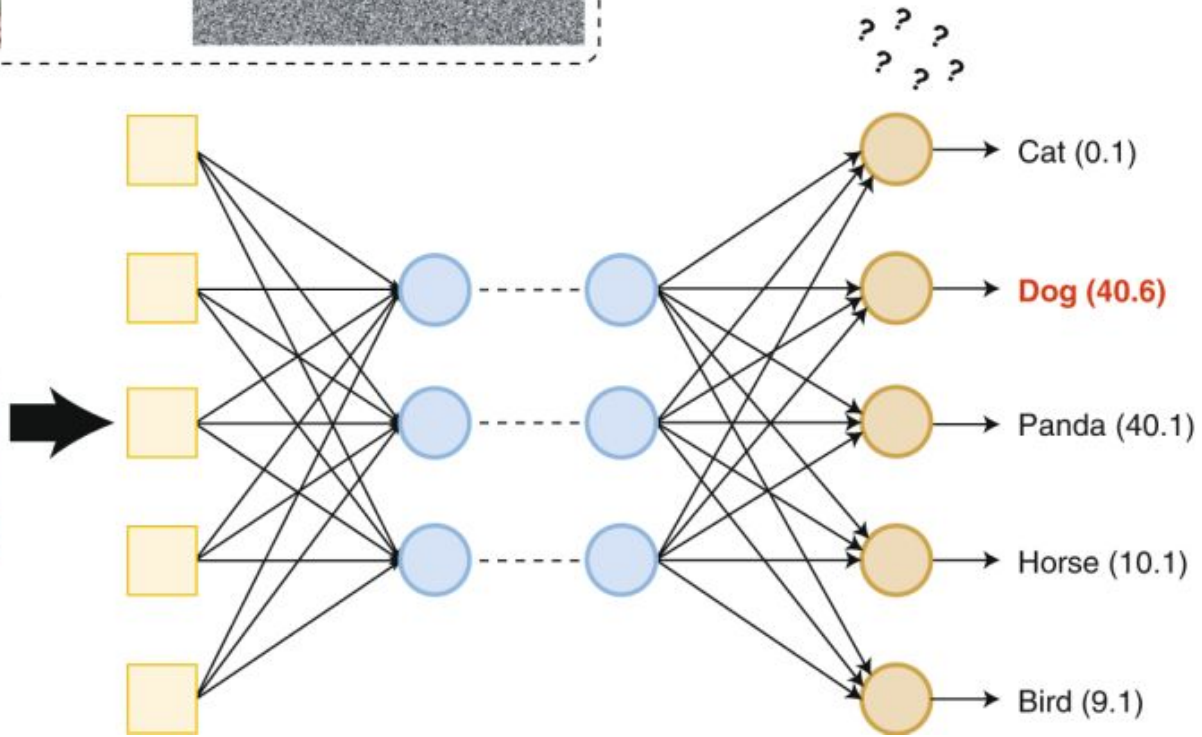
Tempest
[ACADEMY]
Conference

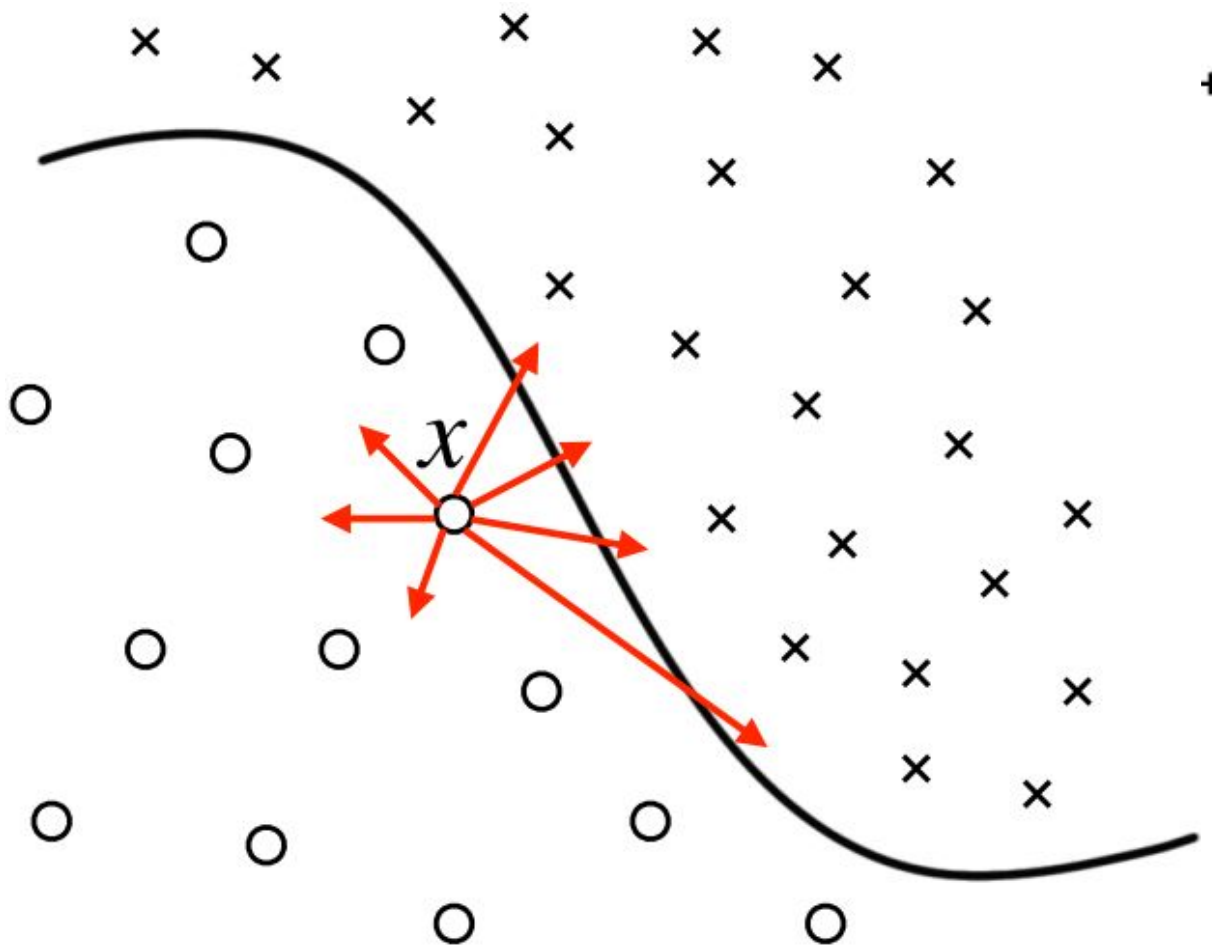
Original



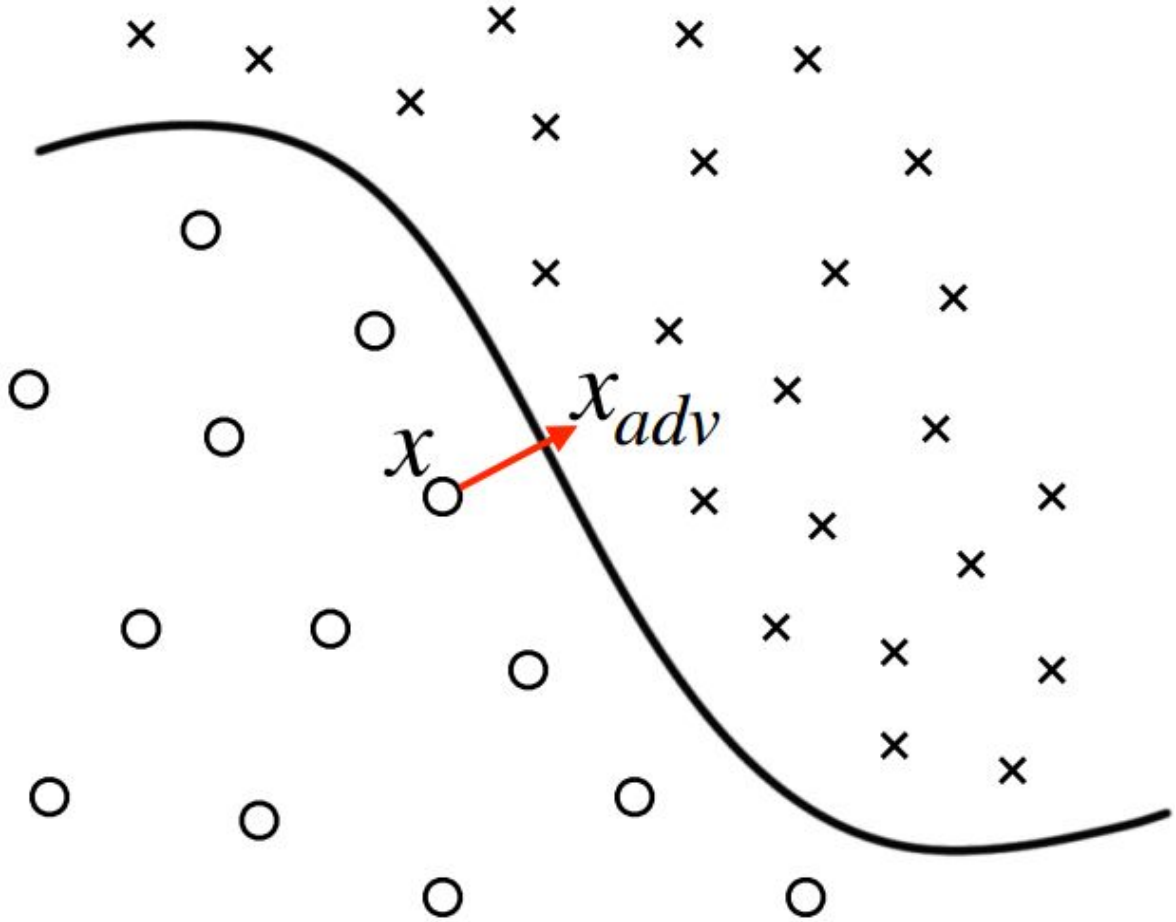
Tempest
ACADEMY
Conference

Adversarial





Tempest
ACADEMY
Conference



$$x_{adv} = x + \delta$$

$$\min||x_{adv} - x|| < \rho$$

$$f(x_{adv}) \neq f(x)$$

Forging adversarial perturbations

Tempest
:ACADEMY:
Conference

IEEE COMMUNICATIONS LETTERS, VOL. 26, NO. 7, JULY 2022

Multi-Objective GAN-Based Adversarial Attack Technique for Modulation Classifiers

Paulo Freitas de Araujo-Filho, Georges Kaddoum, *Senior Member, IEEE*, Mohamed Naili, Emmanuel Thepie Fapi, and Zhongwen Zhu, *Senior Member, IEEE*

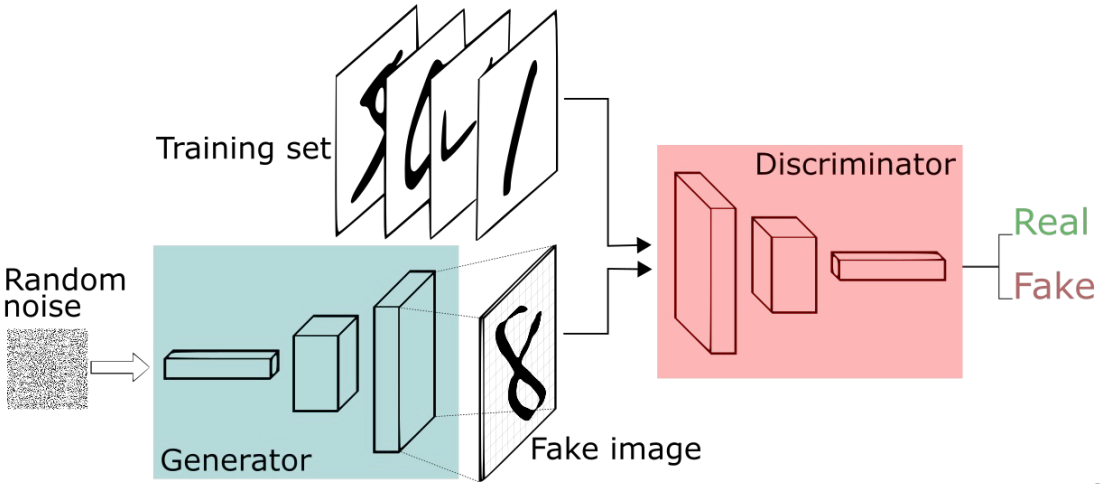
$$f(x_{adv}) \neq f(x)$$

$$\min||x_{adv} - x|| < \rho$$

Generative Adversarial Networks (GANs)

- Gerador G
  - Treinado para produzir amostras sintéticas de dados que sejam reconhecidos para reais
  - Aprende a distribuição de probabilidade de dos dados reais
  - Implicitamente modela o sistema

- Discriminador D
  - Treinado para distinguir as amostras reais daquelas produzidas pelo gerador



$$L_G = -D(G(z))$$

$$L_D = D(G(z)) - D(x)$$

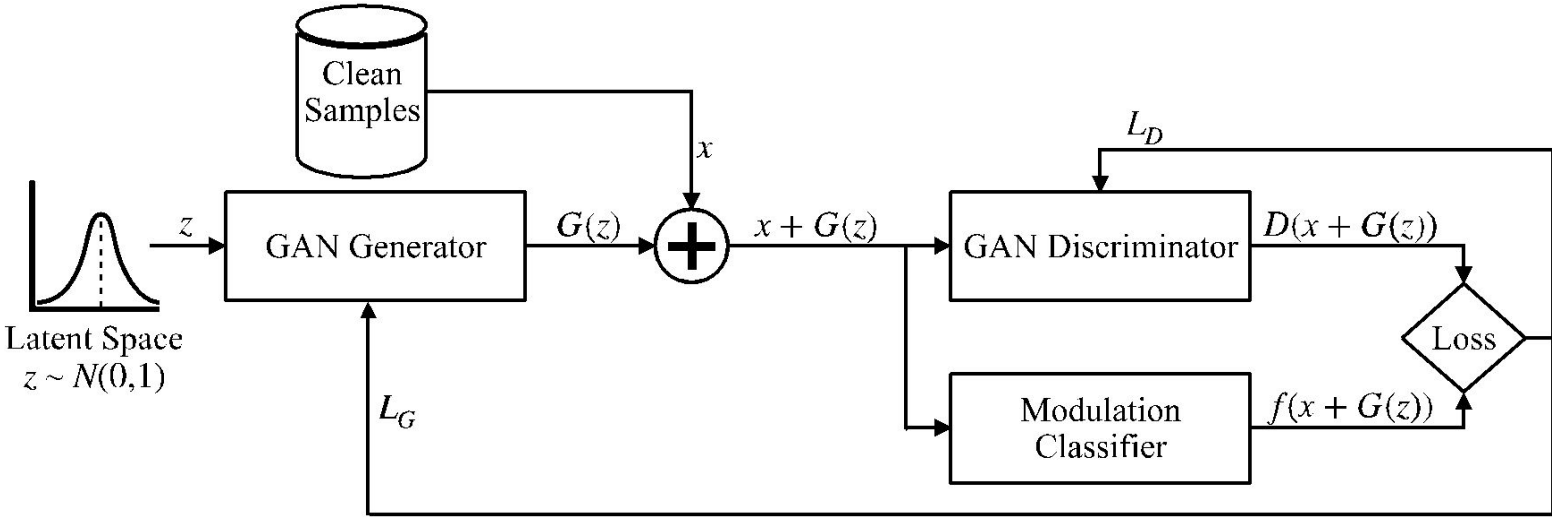
Multi-Objective GAN-Based Adv Attack

Tempest ACADEMY Conference

- Modificamos a estrutura da GAN para que o gerador produza perturbações adversariais

$$\delta = G(z)$$
$$x_{adv} = x + G(z)$$

$$L_G = -D(x + G(z))$$
$$L_D = D\big(x + G(z)\big) - D(x)$$



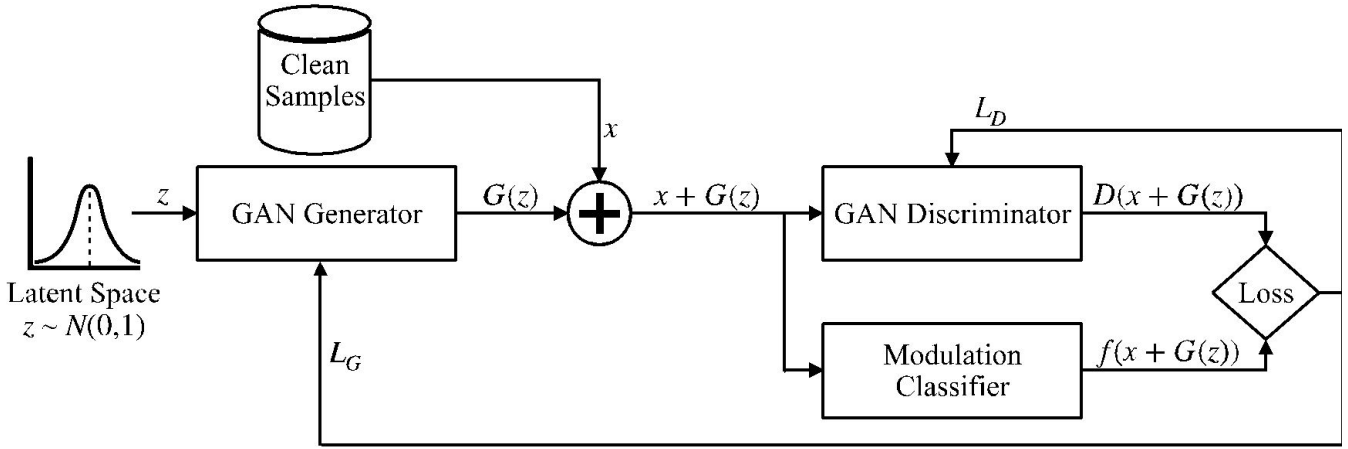
Multi-Objective GAN-Based Adv Attack

Tempest ACADEMY Conference

- Modificamos a estrutura da GAN para que o gerador produza perturbações adversariais

$$L_D = D\big(x + G(z)\big) - D(x)$$

$$L_G = -D(x + G(z))$$

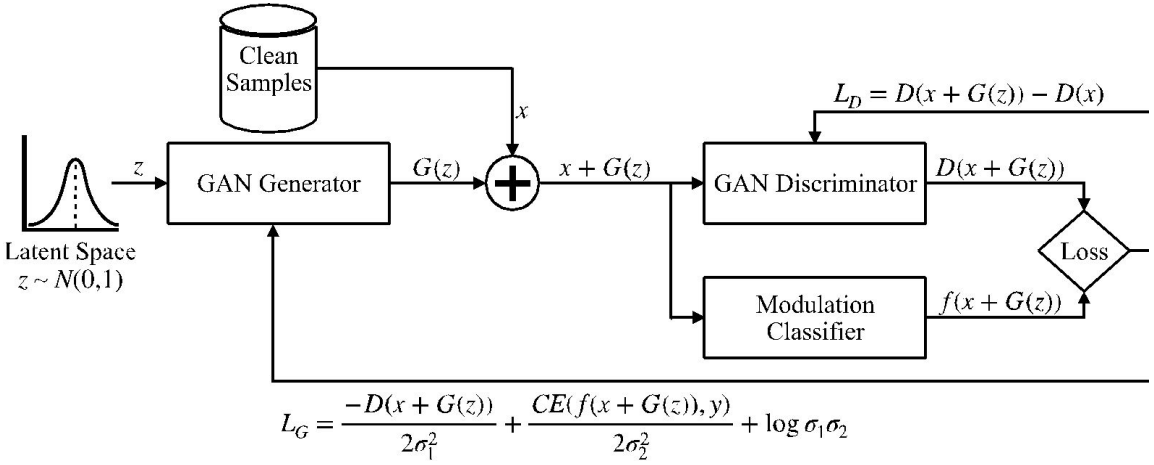


$$L_{G2} = CE(f\big(x + G(z)\big), y) = -\sum_{i=1}^{n} y_i \log(f_i(x + G(z)))$$

Multi-Objective GAN-Based Adv Attack

Tempest ACADEMY Conference

- Modificamos a estrutura da GAN e a combinamos com a Multi-Task Loss



$$L_D = D\big(x + G(z)\big) - D(x)$$

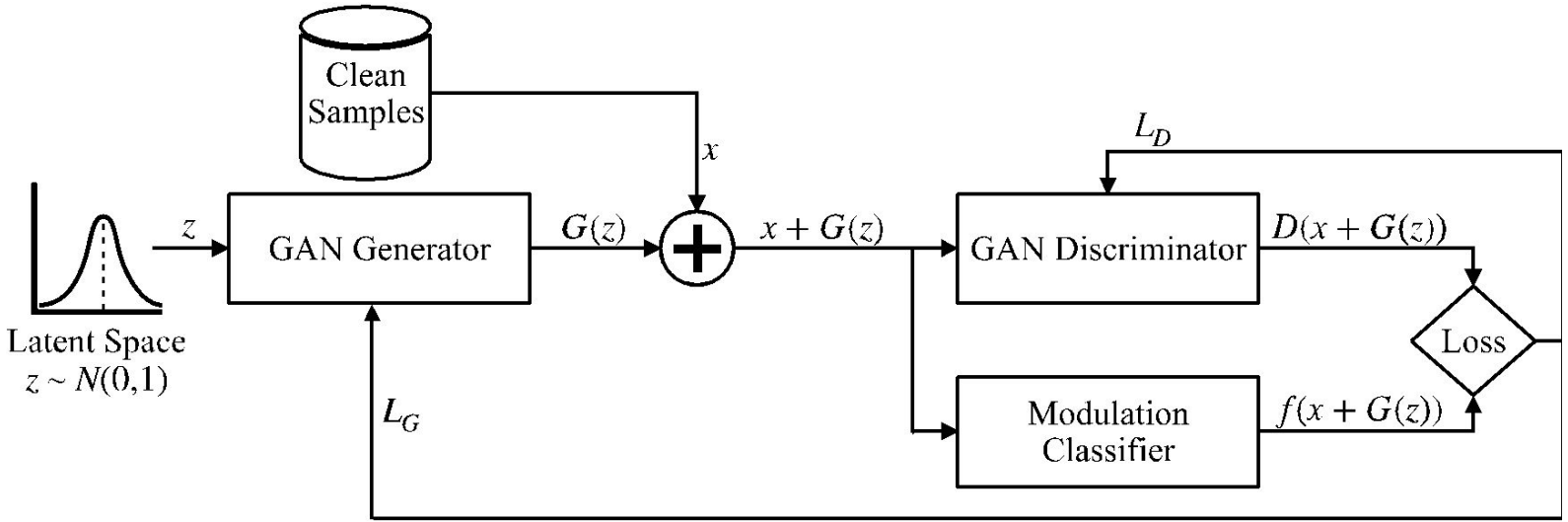
$$L_G = -\frac{D\big(x+G(z)\big)}{2\sigma_1^2} + \frac{CE\big(f\big(x+G(z)\big),y\big)}{2\sigma_2^2} + \log(\sigma_1\sigma_2)$$

Multi-Objective GAN-Based Adv Attack

Tempest ACADEMY Conference

- Modificamos a estrutura da GAN para que o gerador produza perturbações adversariais

$$L_D = D\big(x + G(z)\big) - D(x) \qquad L_G = -\frac{D\big(x+G(z)\big)}{2\sigma_1^2} + \frac{CE\big(f\big(x+G(z)\big),y\big)}{2\sigma_2^2} + \log(\sigma_1\sigma_2)$$

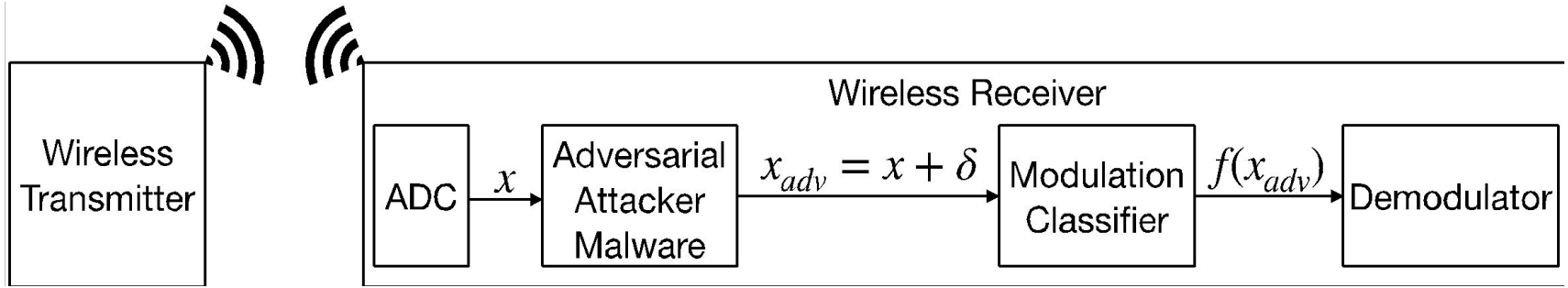


Multi-Objective GAN-Based Adv Attack

Tempest ACADEMY Conference

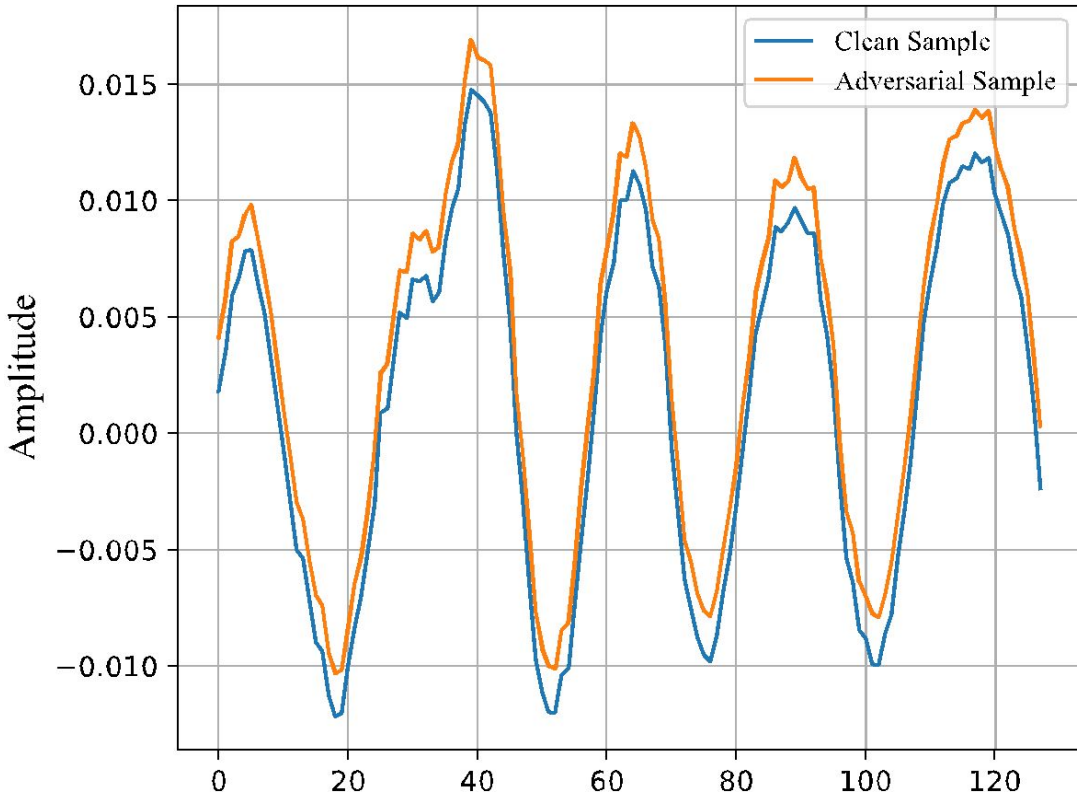
Algorithm 1 Proposed Adversarial Attack Technique

1: Train a GAN according to equations (4) and (5)
2: **for** Each incoming sample x **do**
3: Compute $G(z)$
4: Construct the adversarial sample $x_{adv} = x + G(z)$
5: **end for**



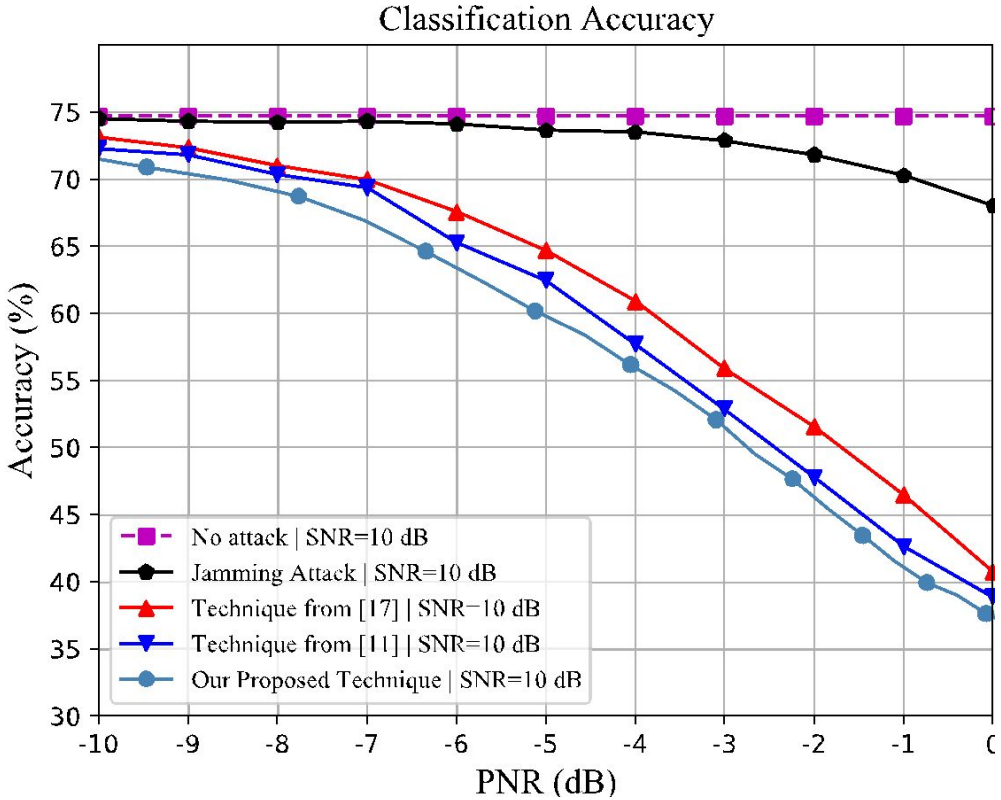
Multi-Objective GAN-Based Adv Attack

Tempest ACADEMY Conference



Multi-Objective GAN-Based Adv Attack

Tempest ACADEMY Conference

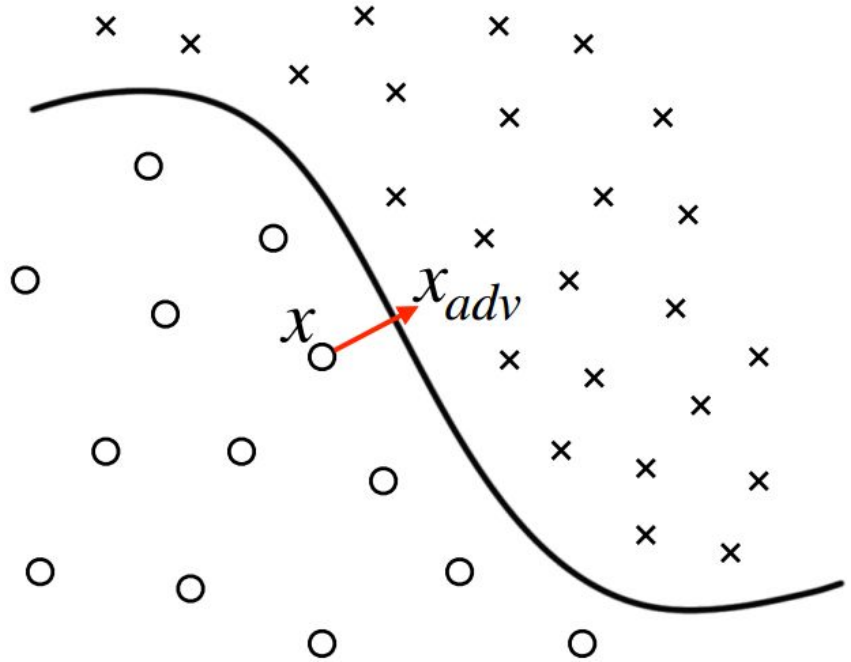


Adversarial Attack Technique	Mean Execution Time per Sample
Technique from [17]	20189 ms
Technique from [11]	234 ms
Our Proposed Technique	0.6980 ms

The Aftermath

Como mitigar?

Tempest
[ACADEMY]
Conference



- Dificultar acesso
- Esconder informações
- Detectar perturbações
- Remover perturbações
- Suavizar fronteiras de decisão
- Combinar modelos

The Defense

Tempest
[ACADEMY]
Conference

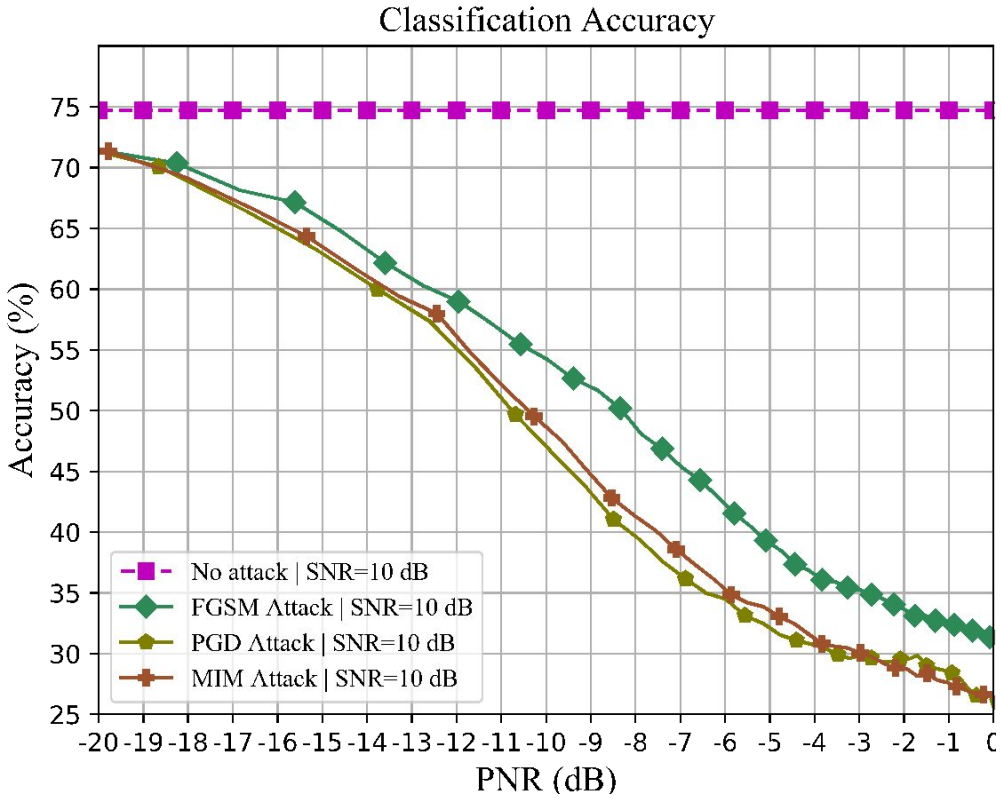
IEEE INTERNET OF THINGS JOURNAL, VOL. 10, NO. 21, 1 NOVEMBER 2023 19153

Defending Wireless Receivers Against Adversarial Attacks on Modulation Classifiers

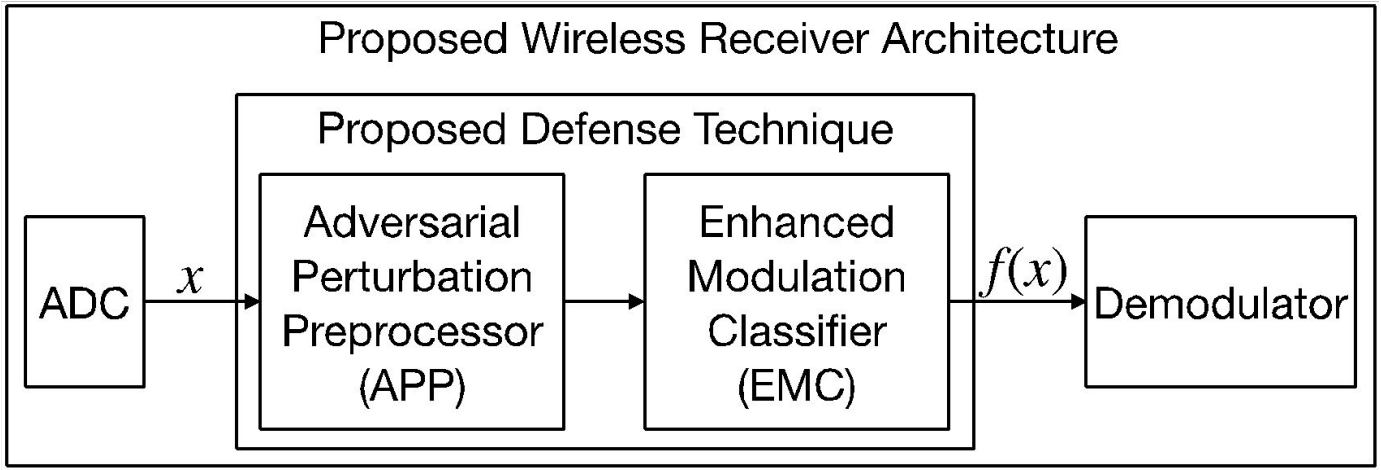
Paulo Freitas de Araujo-Filho, Georges Kaddoum, *Senior Member, IEEE*, Mohamed Chiheb Ben Nasr, Henrique F. Arcoverde, and Divanilson R. Campelo, *Member, IEEE*

The Defense

Tempest
ACADEMY
Conference



The Defense

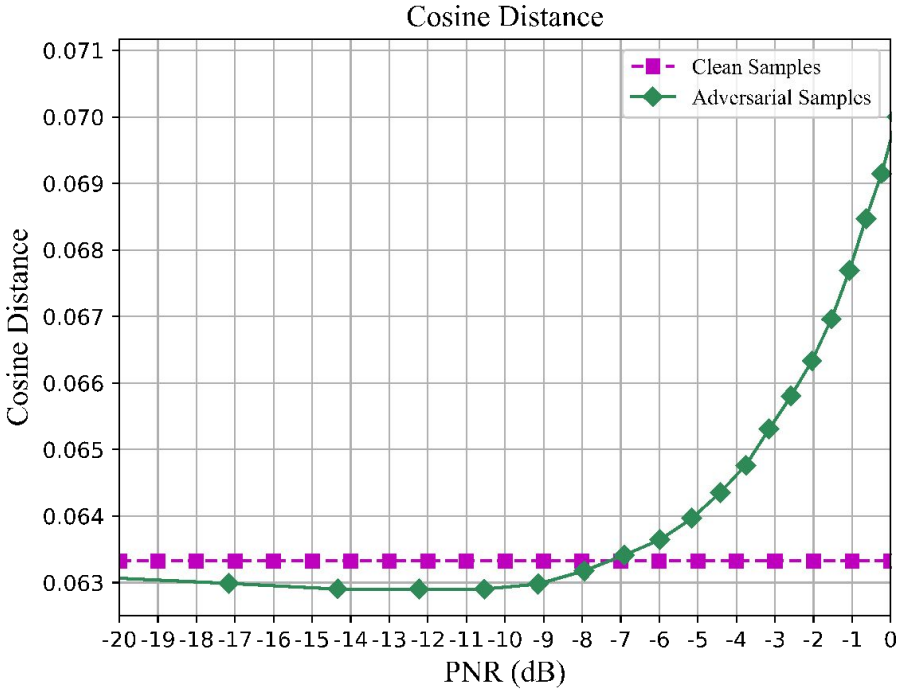

Proposed Wireless Receiver Architecture
Proposed Defense Technique
ADC
x
Adversarial Perturbation Preprocessor (APP)
Enhanced Modulation Classifier (EMC)
$f(x)$
Demodulator

The Defense

Tempest ACADEMY Conference

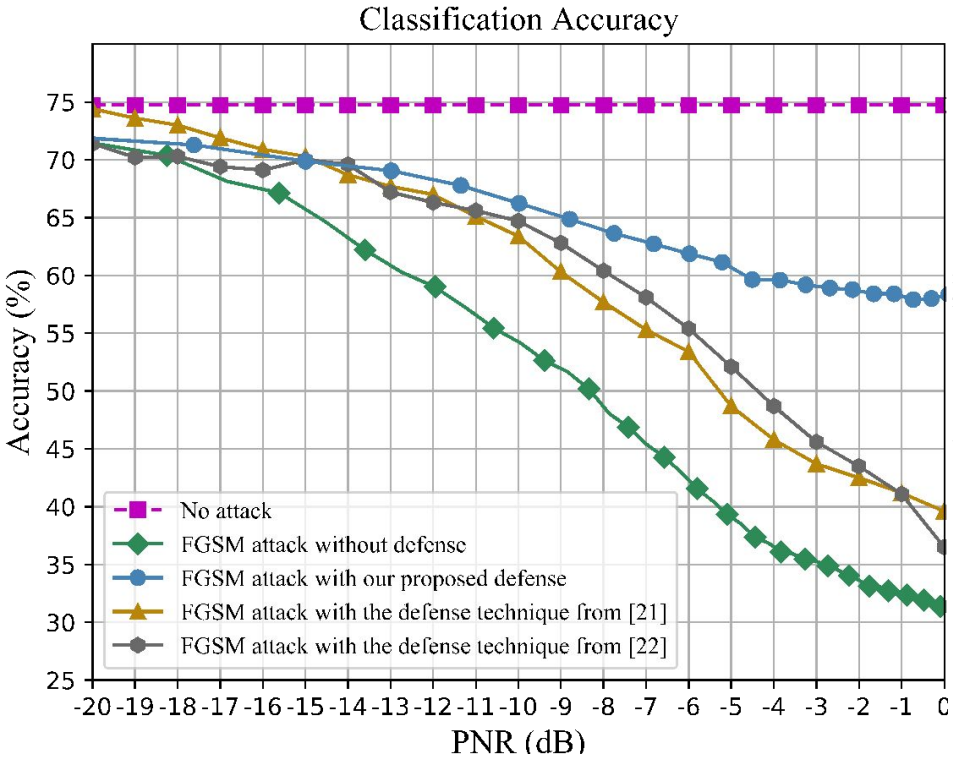
Algorithm 1: Proposed Defense Technique

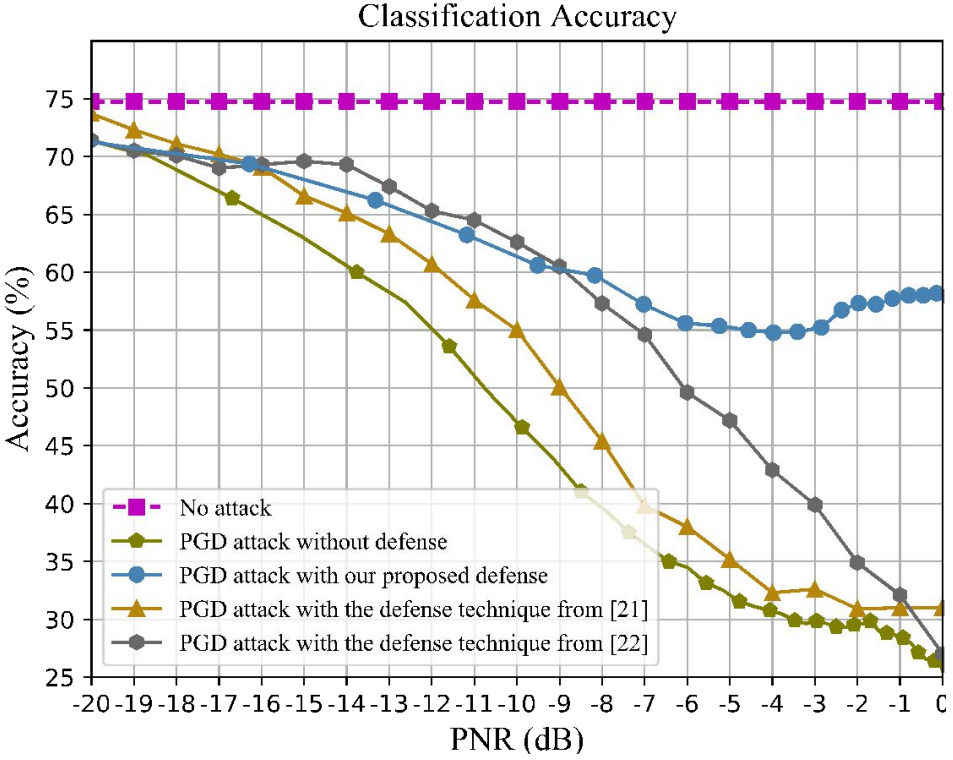
1: Train a DAE with samples tampered with Gaussian noise and adversarial perturbations
2: Train a EMC with samples tampered with Gaussian noise and adversarial perturbations
3: **for** Each incoming sample x_i **do**
4: Compute $x_o = DAE(x_i)$
5: Compute $\beta = CD(x_i, x_o)$
6: **if** $\beta \geq t$ **then**
7: Preprocess data sample $x = x_o$
8: **else**
9: Do not preprocess data sample $x = x_i$
10: **end if**
11: Classify data sample $y = f(x)$
12: **end for**



The Defense

Tempest
[ACADEMY]
Conference





The Future

The Future

Tempest
[ACADEMY]
Conference

- AI brings great changes
- Adversarial attacks are a real problem that can cause devastating consequences to many different applications
- Defense mechanisms exist but cannot yet fully protect ML-based systems.

New and better
defense mechanisms are needed!

Tempest

:ACADEMY:

Conference

2023