Paper Review - TFHM : A Traffic Feature Hiding Scheme Based on Generative Adversarial Networks

Fernando Nakayama

Research Advisor: Dr. Michele Nogueira Lima

Center for Computational Security sCience (CCSC)

Belo Horizonte – Brazil

March 10, 2023

www.ccsc-research.org

Selected Paper

TFHM : A Traffic Feature Hiding Scheme Based on Generative Adversarial Networks

1st Yaya Huang
Cyberspace Institute of Advanced Technology
Guangzhou University
Guangzhou 510006, China
2112006111@e.gzhu.edu.cn

2nd Yixing Chen
Cyberspace Institute of Advanced Technology
Guangzhou University
Guangzhou 510006, China
2112006074@e.gzhu.edu.cn

3rd Yuqiang Zhang
Cyberspace Institute of Advanced Technology
Guangzhou University
Guangzhou 510006, China
2112006277@e.gzhu.edu.cn

4th Jincai Zou
Cyberspace Institute of Advanced Technology
Guangzhou University
Guangzhou 510006, China
2112006304@e.gzhu.edu.cn

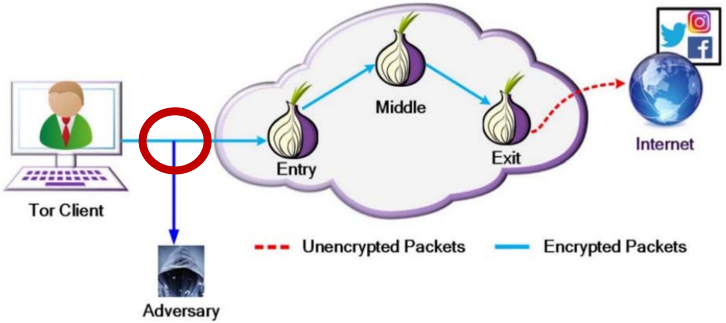
5th Zhihan Tan
Cyberspace Institute of Advanced Technology
Guangzhou University
Guangzhou 510006, China
2112006199@e.gzhu.edu.cn

6th Ning Hu†
Peng Cheng Laboratory
Shenzhen 518000, China
hun@pcl.ac.cn

TFHM : A Traffic Feature Hiding Scheme Based on Generative Adversarial Networks

Problem

- Privacy leakage (encrypted traffic)
- Difficult to implement traffic features hiding mechanisms



UFPR UNIVERSIDADE FEDERAL DO PARANÁ
UFMG

TFHM : A Traffic Feature Hiding Scheme Based on Generative Adversarial Networks

Motivation

- Current defense schemes lack dynamics
- Most schemes lose defense ability

Contribution

- A dynamic traffic feature hiding technology for traffic analysis (TFHM)

CCSC

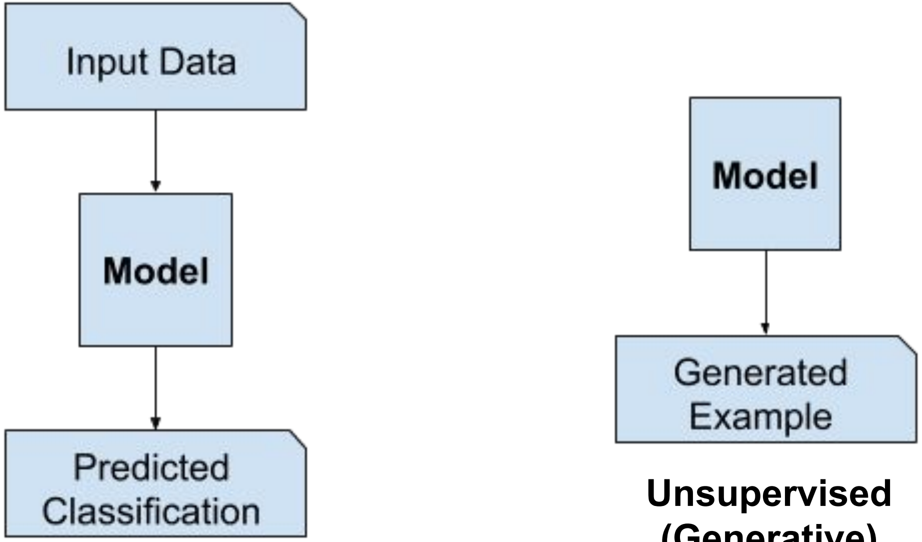
Background

- Adversarial Examples (AML)
- Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GAN)

- **GAN**
  - Generative modeling
  - Unsupervised learning
  - Based on minimax game-theory
  - Generate or output new examples

Generative Adversarial Network (GAN) - Explained

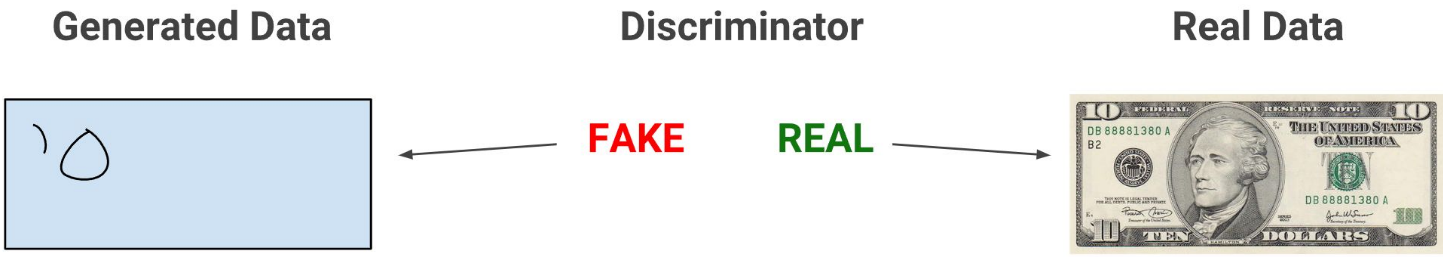


Supervised (Discriminative)

Unsupervised (Generative)

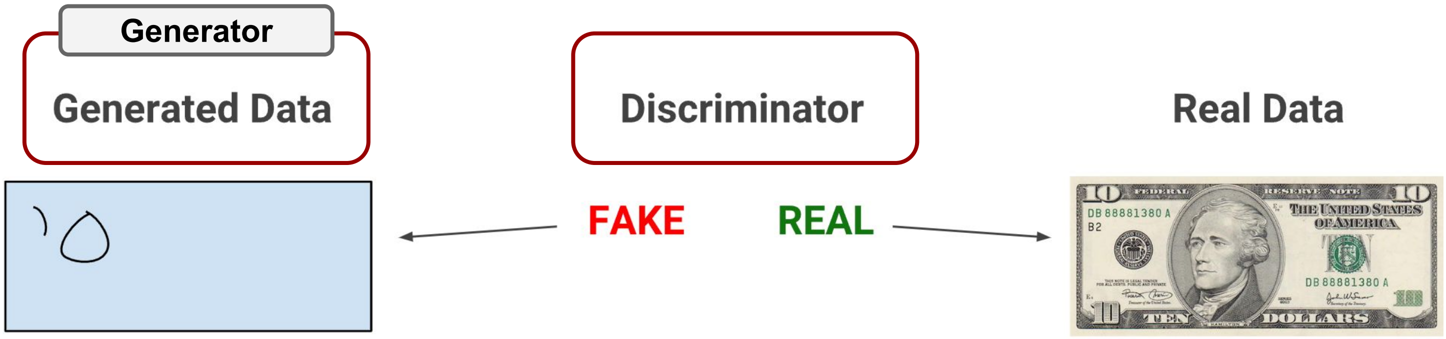
Generative Adversarial Network (GAN)

- Overview of GAN Structure



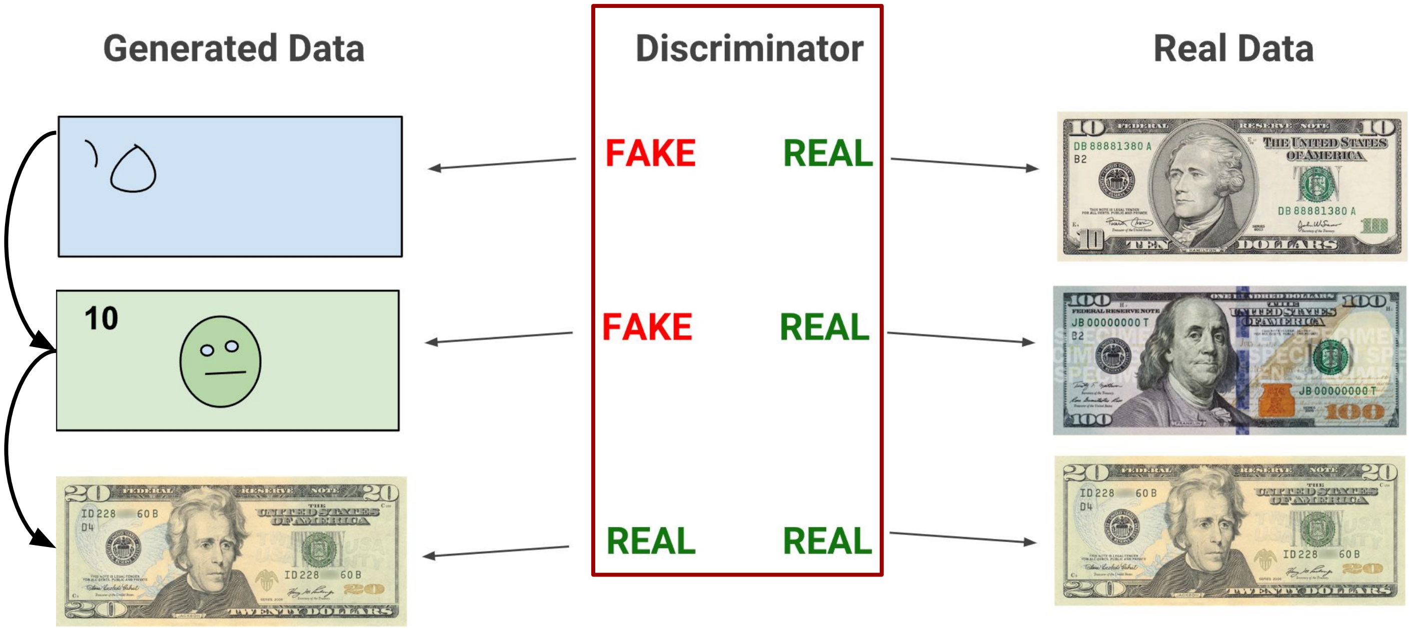
https://developers.google.com/machine-learning/gan/gan_structure

- Overview of GAN Structure



https://developers.google.com/machine-learning/gan/gan_structure

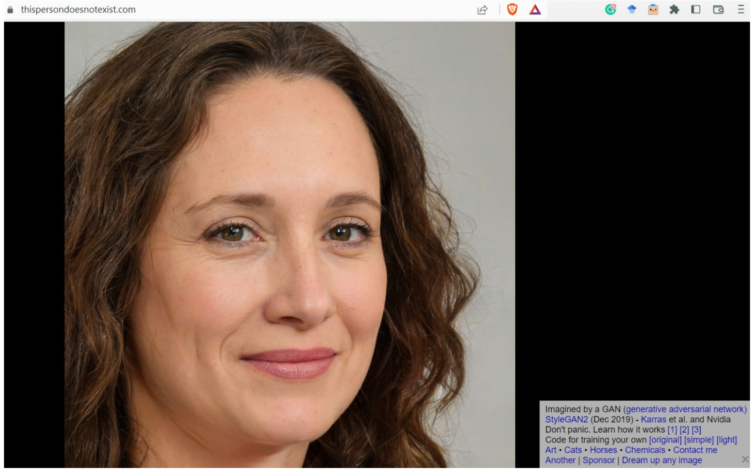
Generative Adversarial Networks (GAN)



https://developers.google.com/machine-learning/gan/gan_structure

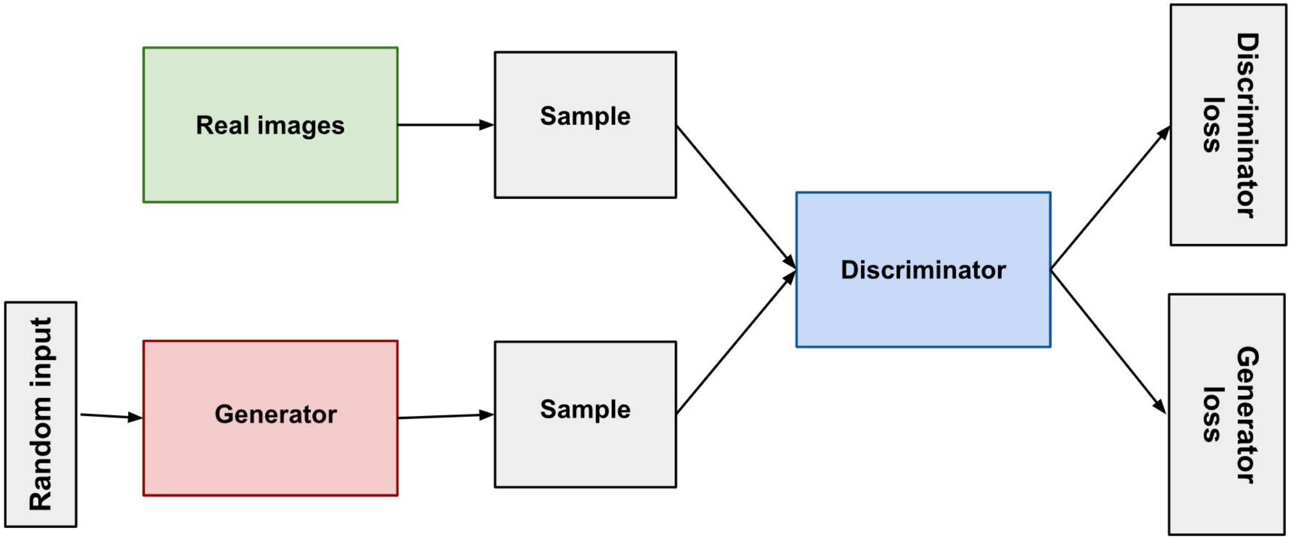
Generative Adversarial Networks (GAN)

Brundage, Miles, et al. "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation." *arXiv preprint arXiv:1802.07228* (2018).



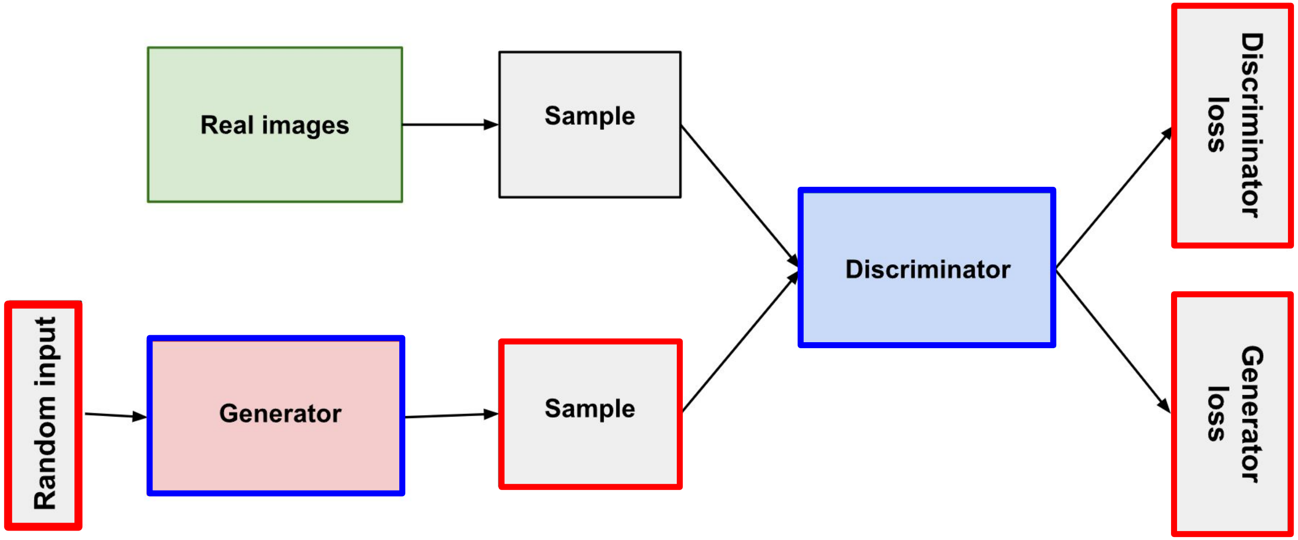
UFPR UFMG

Generative Adversarial Network (GAN)



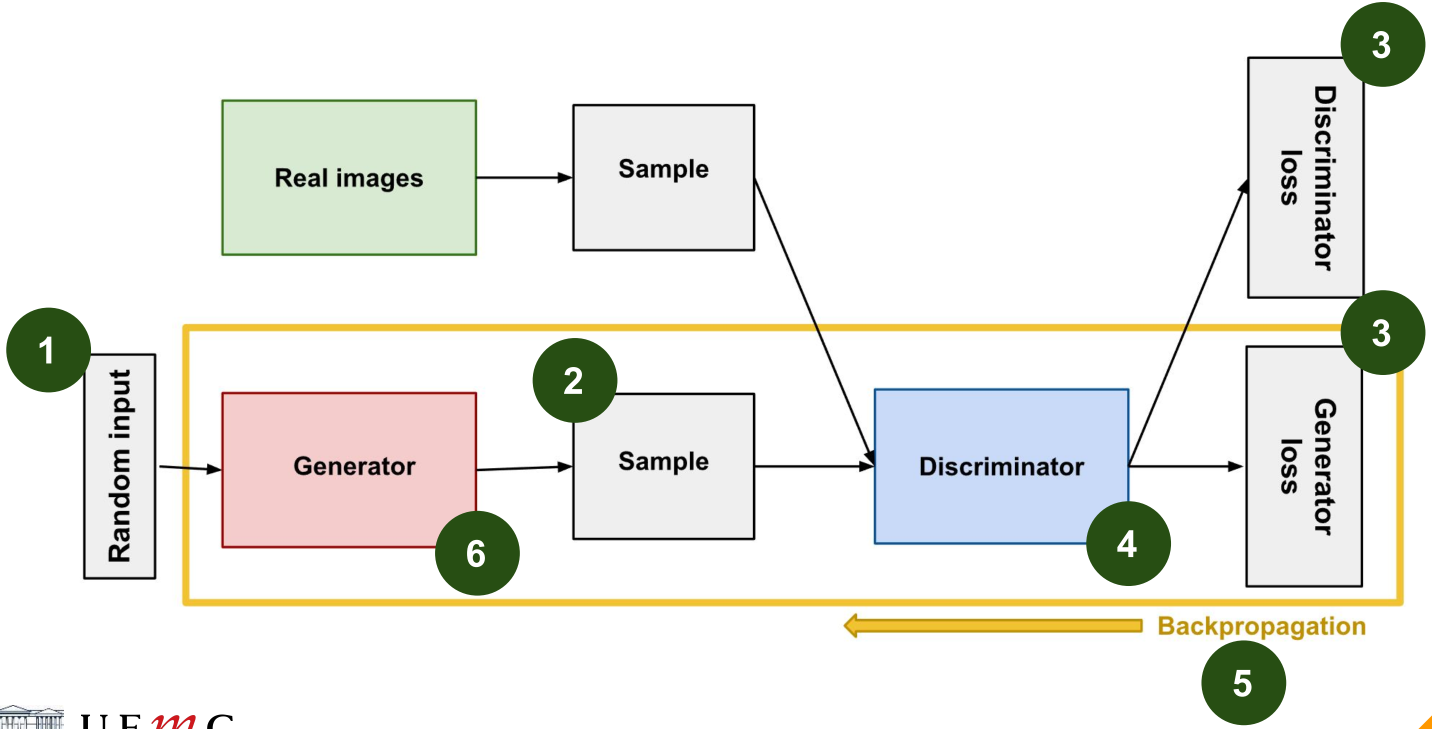
https://developers.google.com/machine-learning/gan/gan_structure

Generative Adversarial Network (GAN)



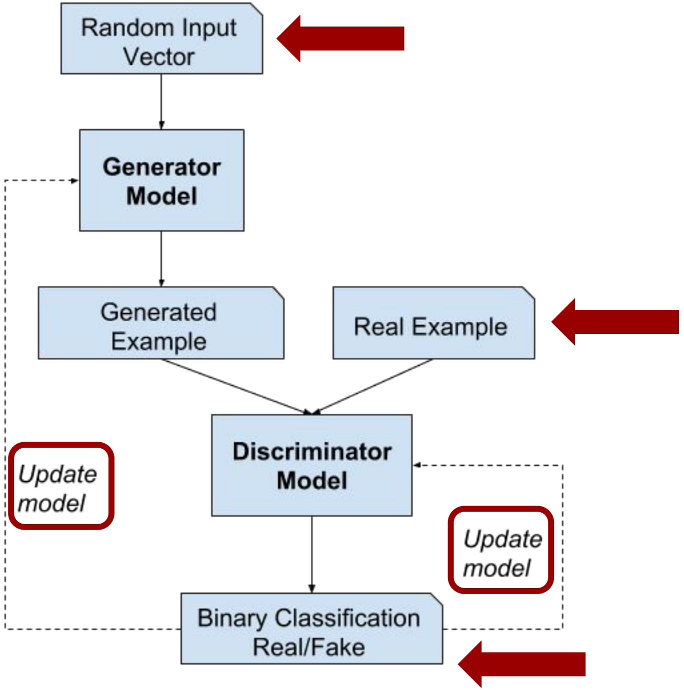
https://developers.google.com/machine-learning/gan/gan_structure

Generative Adversarial Network (GAN)



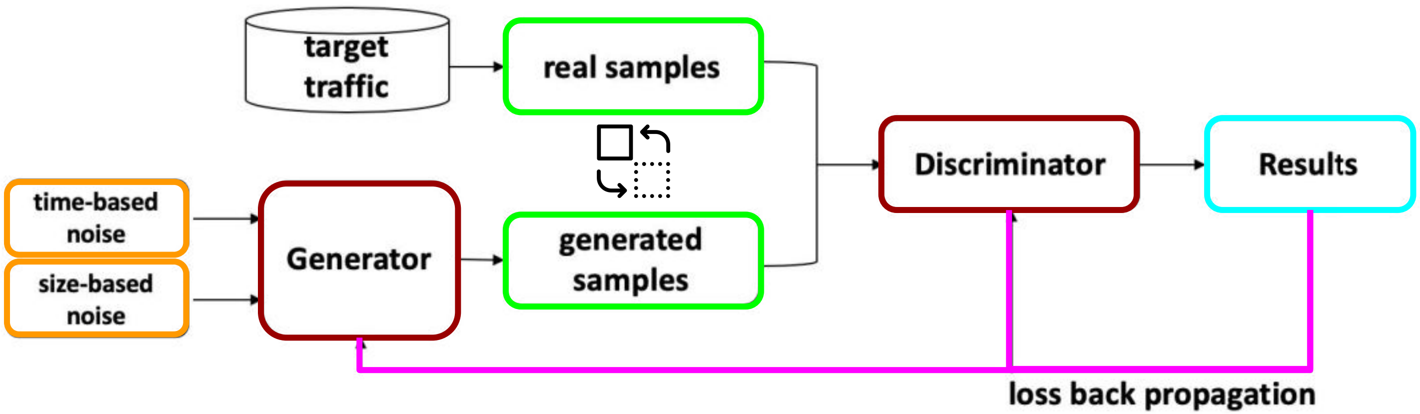
https://developers.google.com/machine-learning/gan/gan_structure

Generative Adversarial Network (GAN)

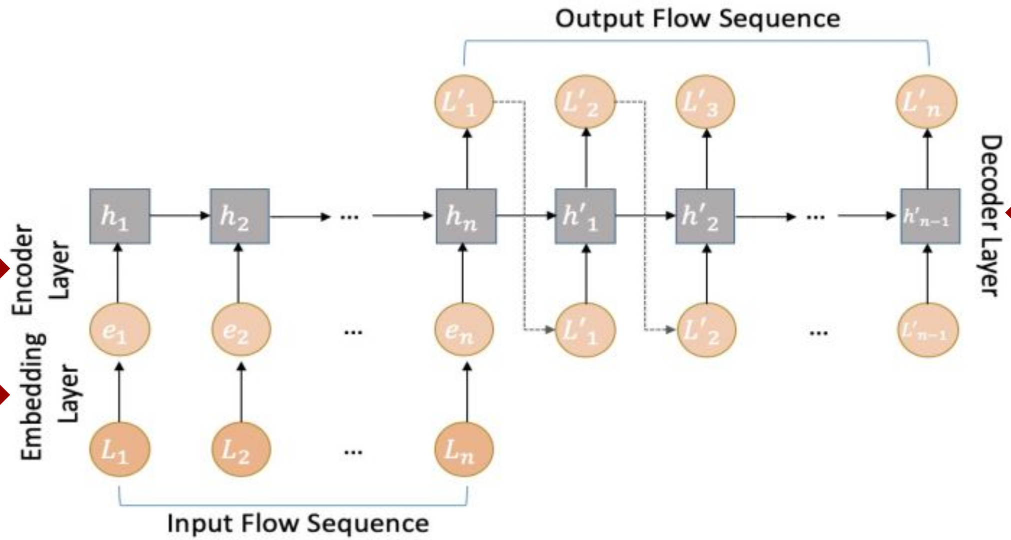


A Gentle Introduction to Generative Adversarial Networks (GANs) - Jason Brownlee

Generative Adversarial Networks (GAN)



Generative Adversarial Network (GAN)



Generator model structure

Seq2Seq model

3 layer (Embedding, encoder, decoder)

Discriminator model structure

Multilayer fully connected network

SIGMOD classification

Generative Adversarial Network (GAN)

Model solution

Minimax game-theory

Algorithm 1 Traffic feature hiding based on GAN

Input: Traffic features from the training set, where is one flow feature including packet sizes sequence and IPDs sequence.

Output: A trained traffic feature generator.

function TRAINALGORITHM
  Intialize a generator G with parameters and a discriminator D with parameters, maximum number of iterations $MAXEPOCH$.
  Current iteration number, $epoch \leftarrow 1$.
  while $epoch < MAXEPOCH$ **do**
    for $i = 0 \rightarrow step$ **do**
      $x \leftarrow$ a batch of m training samples from.
      $z \leftarrow$ a batch of m generated samples from random noise
      $z' \leftarrow$ sample data generated by the generation model based on seq-2-seq
      Update the generator with Adam algorithm by descending the generator's loss:
      $J_G = -E_m[D_\omega G_\theta(z))]$
      Use the discriminator network to distinguish the generated samples and real samples.
      Update the discriminator with Adam algorithm by descending the discriminator's loss:
      $J_D = E_m[D_\omega G_\theta(z)) - D_\omega(x) + \lambda(||)]$
    end for
  end while
  return
end function

Generative Adversarial Networks (GAN)

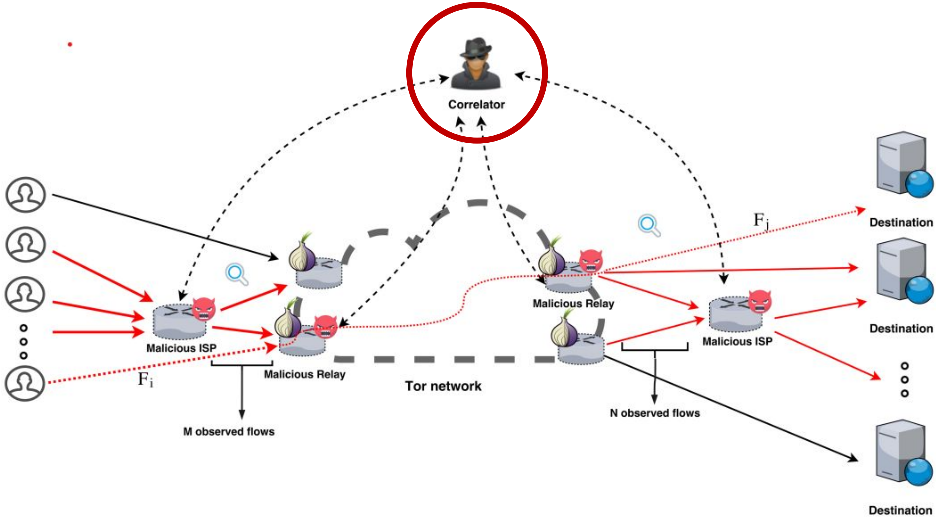
Threat model and dataset

- **DeepCorr**

Inter-packet delay

Packet size

https://github.com/SPIN-UMass/DeepCorr



Requirements

```
tensorflow
tqdm
pickle
numpy
```

Nasr, Milad, Alireza Bahramali, and Amir Houmansadr. "Deepcorr: Strong flow correlation attacks on tor using deep learning." *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. 2018.

Generative Adversarial Networks (GAN)
Results

ACCURACY OF ATTACKS WITH WITHOUT TFHM

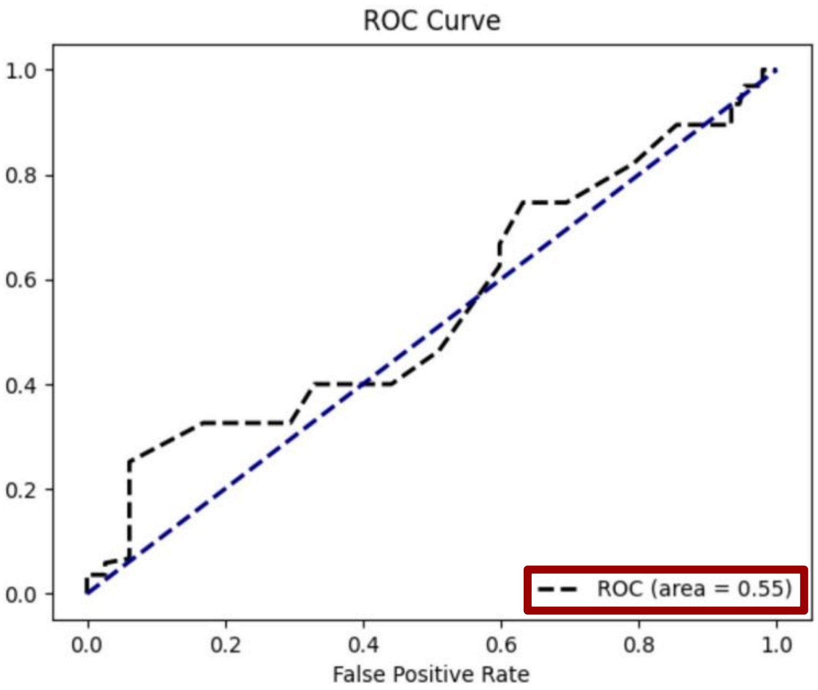
ATTACKS	Acc with TFHM	Acc without TFHM
Naive Bayes	0.593	0.911
Decision Tree	0.612	0.936
SVM	0.569	0.895
DeepCorr	0.749	0.951

DEFENSE MODEL COMPARISON

Model Name	DeepCorr	Full-Connection Networks	TFHM
Accuracy	0.951	0.843	0.749

Yuan S. Research on the generation technology of encrypted traffic based on generative adversarial nets. [D]. Hubei China : Hubei University, 2020.

Generative Adversarial Networks (GAN)
Results



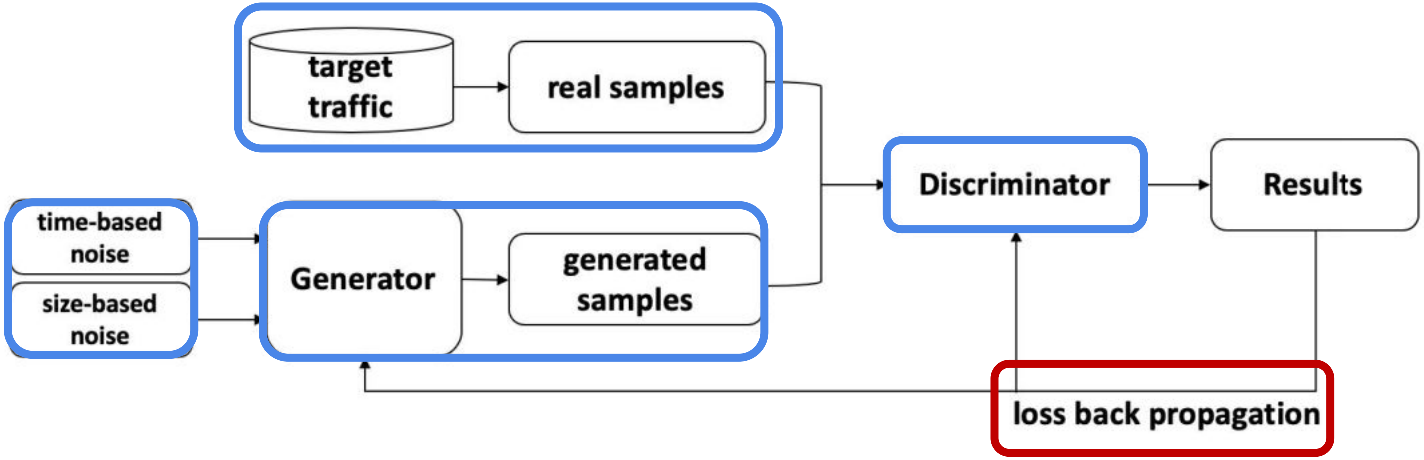
UFPR UNIVERSIDADE FEDERAL DO PARANÁ
UFMG

Generative Adversarial Networks (GAN)
Problems

Limitations

- Limited features contribute to attackers
- Discriminator optimization
- Hybrid models

Generative Adversarial Networks Framework



Generative Adversarial Networks (GAN)
Possibilities

- Deep-learning-based unsupervised framework
- Generated data similar to real data
  - Full datasets
  - Balance existing datasets

Generative Adversarial Networks (GAN)
Problems

- Difficult to train
- Unfeasible for real-time or near-real-time applications
- New threats probably require new models
- Does not tackle the threats identification problem

Problems regarding the Project

fernandonakayama@ufpr.br