Adversarial Attacks in Artificial Intelligence And Solutions

Slides Preparation: Mohammad Amin Dadgar

Table Of Contents

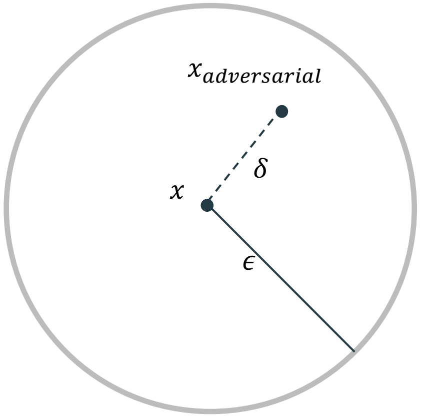
Adversarial Attacks & Defense

- Adversarial Attacks
  - Using adversarial examples fool the model
  - Adversarial examples
    - Add *imperceptible noise* to the original data
    - In order to change the model output
  - First introduced for Computer Vision applications
  - Then extended for video and speech applications
- Adversarial Defense
  - Propose a strategy to Defend the attacks

Adversarial Attacks

Adversarial Attacks - Categorization

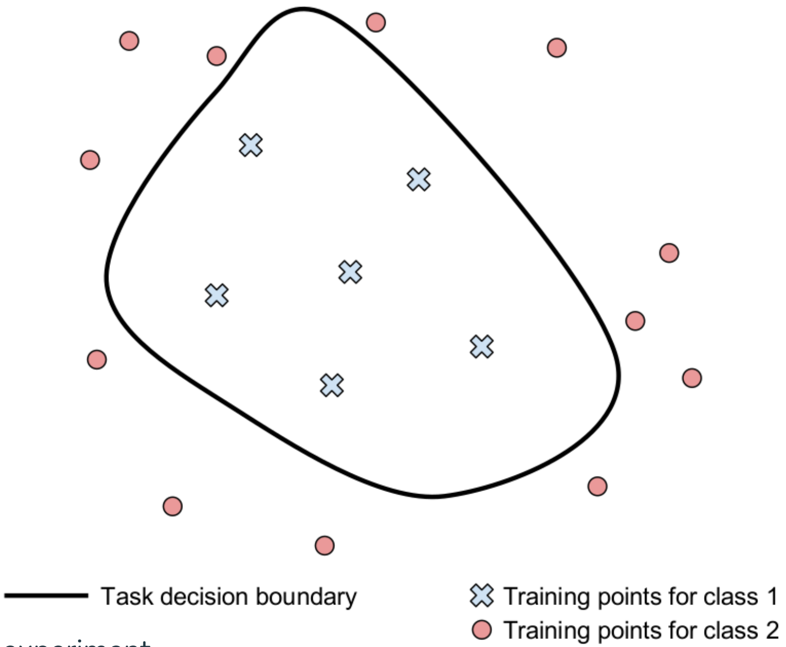
- Adversarial examples are assumed to be produced by
  - $x_{adversarial} = x_{original} + \delta$
  - With a distance constraint to enforce imperceptibility
    - $D(x_{original}, x_{adversarial}) < \epsilon$
- Categorization based on the information of the model
  - White-box attacks
  - Black-box attacks
  - Gray-box attacks



[1] Study of Pre-Processing Defenses Against Adversarial Attacks on State-of-the-Art Speaker Recognition Systems, Sonal Joshi, 2021 IEEE transactions on information forensics and security

Adversarial Attacks – Visual Explanations

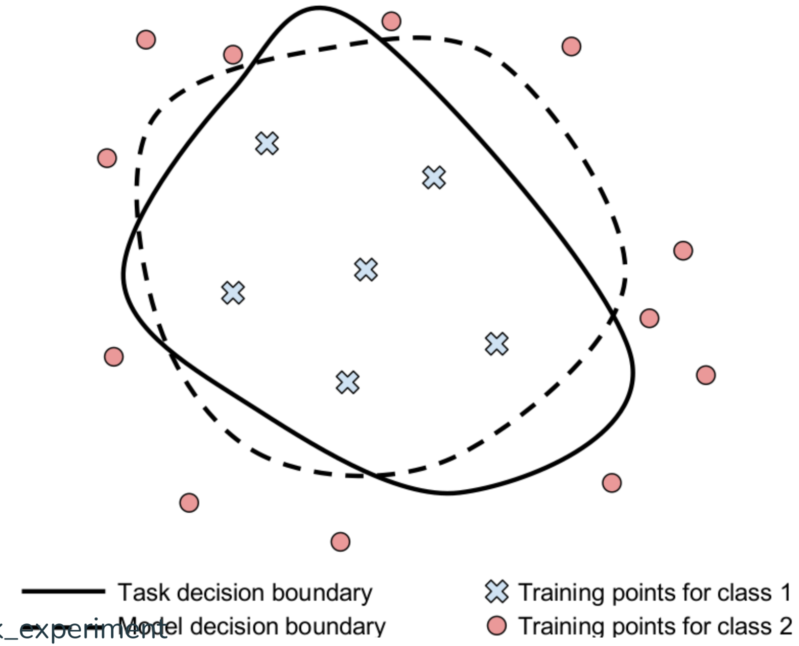
- Dataset And boundaries



Ref: https://github.com/osm3000/adversarial_attack_experiment

Adversarial Attacks – Visual Explanations

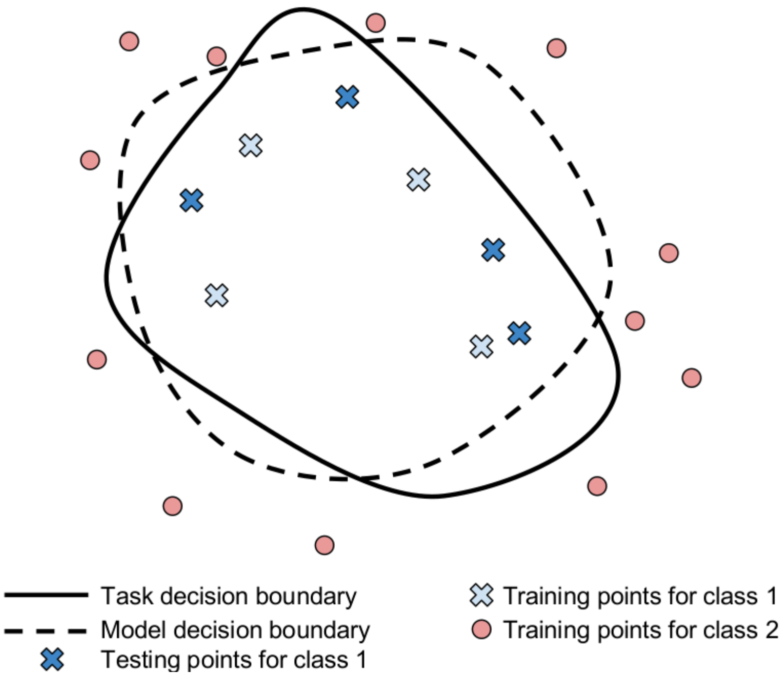
- Training phase



Ref: https://github.com/osm3000/adversarial_attack_experiment

Adversarial Attacks – Visual Explanations

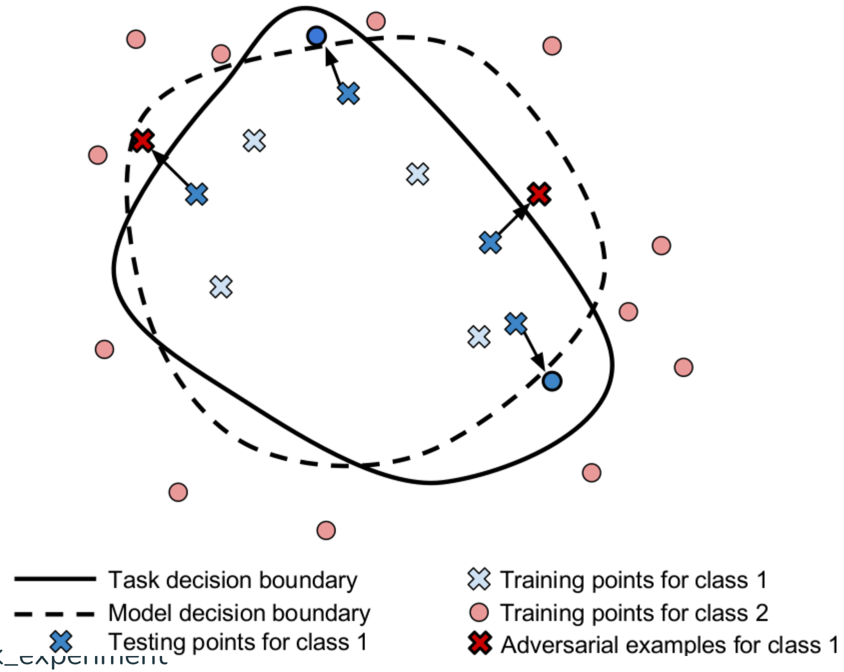
- Testing phase
  - Test Accuracy is 100%



Ref: https://github.com/osm3000/adversarial_attack_ex

Adversarial Attacks – Visual Explanations

- Adversarial examples



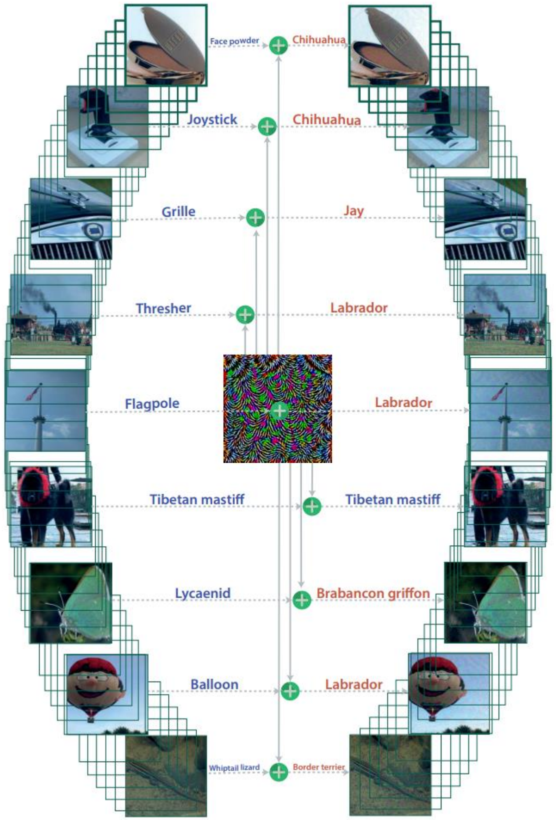
Ref: https://github.com/osm3000/adversarial_attack_experiment

Adversarial Attacks - Methods

- Based on input gradients
    - Fast Gradient Sign Method (FGSM)
        - Adds ϵ or $-\epsilon$ to the data points based on sign of the input gradient
        - Specifically $x_{adversarial} = x_{original} + \epsilon\ sign\left(\nabla_x L\left(g(x), y_{original}\right)\right)$
    - Basic Iterative Method (BIM a.k.a. Iterative FGSM)
        - Generate Iteratively with smaller steps than FGSM
    - Projected Gradient Descent (PGD)
        - Iterative as BIM
        - Considers all norm distances (1, 2, ..., ∞ norms)
        - Randomly starts from an adversarial point in ϵ ball
            - Instead of starting from the original x

Adversarial Attacks - Methods

- Universal attacks [1]
  - Find One universal perturbation in which it can change the most data predictions



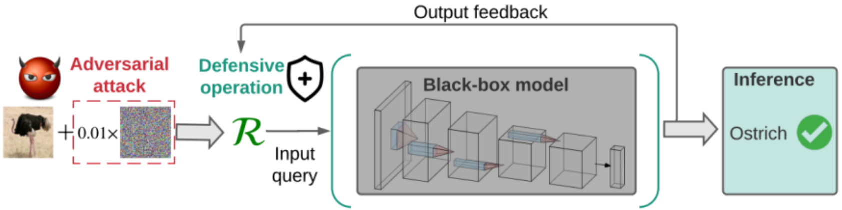
[1] Universal adversarial perturbations, Seyed-Mohsen Moosavi-Dezfooli, CVPR 2017

Adversarial Defense

Adversarial Defenses - Categorization

- There are two strategy [1]
  - Proactive defenses
    - Re-train the model with adversarial examples
  - Reactive defenses
    - Adds a block to the network
      - Reject the adversarial samples
      - Reconstruct the adversarial samples

Reactive Defense [2]



[1] Study of Pre-Processing Defenses Against Adversarial Attacks on State-of-the-Art Speaker Recognition Systems, Sonal Joshi, 2021 IEEE transactions on information forensics and security
[2] How to Robustify Black-Box ML Models? A Zeroth-Order Optimization Perspective, Yimeng Zhang, 2022 ICLR Spotlight

Adversarial Defenses – Methods

- Adversarial Training
  - Proactive Defense
- GAN networks
  - Proactive
    - Re-train the model using the produced sample
  - Reactive
    - Gives the nearest produced sample to the original input
- VAE Defense
  - Both Proactive and Reactive

[1] Study of Pre-Processing Defenses Against Adversarial Attacks on State-of-the-Art Speaker Recognition Systems, Sonal Joshi, 2021 IEEE transactions on information forensics and security

Experiments

Experiments

- VoxCeleb2 dataset
  - 6114 speakers
  - Training data was augmented 6 times
    - With MUSAN corpus noise dataset
    - Impulse responses from the RIR dataset
- Networks
  - ResNet34
  - EfficientNet-b0/b4
  - Transformer-Encoder
  - ThinResNet34
  - The fusion of above

Experiments – Undefended Baselines

TABLE I

IDENTIFICATION ACCURACY (%) FOR SEVERAL UNDEFENDED X-VECTOR ARCHITECTURES UNDER ADVERSARIAL ATTACKS

Architecture	Clean	FGSM Attack					BIM Attack					Universal	CW
L_∞		0.0001	0.001	0.01	0.1	0.2	0.0001	0.001	0.01	0.1	0.2	0.3	-
1. ResNet34	100.0	99.1	95.8	95.6	93.3	87.2	92.2	14.8	0.0	0.0	0.0	100.0	1.3
2. EfficientNet-b0	100.0	99.2	95.6	93.0	93.6	88.1	96.9	27.7	0.0	0.0	0.0	100.0	0.8
3. EfficientNet-b4	100.0	99.5	95.8	92.3	93.1	88.8	98.1	30.5	0.0	0.0	0.0	100.0	0.0
4. Transformer	99.5	96.3	80.6	76.4	49.5	32.1	81.9	20.3	0.2	0.0	0.0	99.9	1.9
5. ThinResNet34	100.0	98.0	91.1	89.2	85.6	74.5	88.0	2.2	0.0	0.0	0.0	100.0	1.1
Fusion 2+4+5	100.0	99.8	97.5	97.0	88.4	78.0	98.9	66.4	16.1	0.0	0.2	100.0	49.1

[1] Study of Pre-Processing Defenses Against Adversarial Attacks on State-of-the-Art Speaker Recognition Systems, Sonal Joshi, 2021 IEEE transactions on information forensics and security

Experiments - Attacks

- Adversarial Robustness Toolkit From IBM
  - *Trusted-AI* GitHub account
- FGSM, BIM (I-FGSM)
  - L_∞
  - norms ϵ between 0.0001 and 0.2
- BIM (I-FGSM)
  - $\alpha = \epsilon/5$, with iterations 7, 50, 100
- PGD
  - Learning rate $\alpha = \epsilon/5$, 10 random restarts, with iterations 50 and 100
- CW-L2
  - Confidence $\kappa = 0$ and learning rate 0.001,
  - 10 iterations inner loop and maximum 10 iterations outer loop
- Universal Perturbation
  - Transfer black-box attacks from a SincNet Model

Experiments – Defense Strategies

TABLE VII

SUMMARY OF IDENTIFICATION ACCURACY (%) OF ALL DEFENSES WITH THEIR BEST SETTING. NOTE: SMOOTHING $\sigma = 0.2$, PGD/FGSM ADVTR $\varepsilon = \mathcal{U}(0, 0.01)$, PWG MODELS IS TRAINED ON VOXCELEB. FOR ADAPTIVE ATTACKS, PWG/VAE DEFENSES ARE EITHER APPROXIMATED (BPDA) OR END-TO-END DIFFERENTIABLE (E2ED)

Defense	Clean	FGSM Attack					BIM Attack					Universal	CW
L_∞	-	0.0001	0.001	0.01	0.1	0.2	0.0001	0.001	0.01	0.1	0.2	0.3	-
No defense	100.0	96.9	90.0	92.3	93.4	91.1	83.4	2.3	0.0	0.0	0.0	100.0	1.4
PGD AdvTr	75.5	76.4	75.3	59.8	25.0	18.1	75.8	72.7	39.4	8.9	8.4	87.9	30.3
FGSM AdvTr	89.1	89.2	88.3	89.5	63.6	49.8	89.1	77.0	24.5	7.2	7.0	95.9	32.3
Smoothing	98.0	98.3	98.4	97.0	64.4	44.1	97.2	97.8	97.7	18.9	2.0	98.7	96.9
DefenseGAN	96.3	91.6	84.2	81.9	49.4	23.3	85.0	23.6	2.8	1.4	1.6	96.9	60.9
VAE *BPDA*	98.9	98.4	98.1	94.2	91.4	84.8	94.7	67.2	12.7	0.9	1.1	99.9	56.1
VAE *E2ED*	99.4	96.9	94.1	94.2	91.4	84.8	92.3	35.3	1.3	0.9	0.5	99.8	50.2
Smoothing before VAE *BPDA*	95.2	95.8	95.5	94.7	79.5	51.4	96.6	95.2	94.5	63.1	9.8	97.5	95.5
Smoothing before VAE *E2ED*	95.2	95.8	96.1	95.2	65.6	50.0	96.1	95.8	94.2	19.1	2.7	97.4	95.8
PWG *BPDA*	99.5	99.5	99.7	99.1	86.6	77.2	99.4	99.7	99.5	97.2	92.3	99.9	98.8
PWG *E2ED*	97.0	98.8	95.8	93.6	83.0	62.3	94.7	36.4	0.8	0.8	0.8	99.8	37.5
Smoothing before PWG *BPDA*	95.6	95.2	96.3	95.8	93.0	74.2	94.8	94.8	96.9	95.5	93.4	97.5	95.2
Smoothing before PWG *E2ED*	95.8	94.7	95.6	94.4	88.9	60.6	95.2	93.3	86.7	14.4	3.1	97.4	92.8

Experiments – Defense Strategies

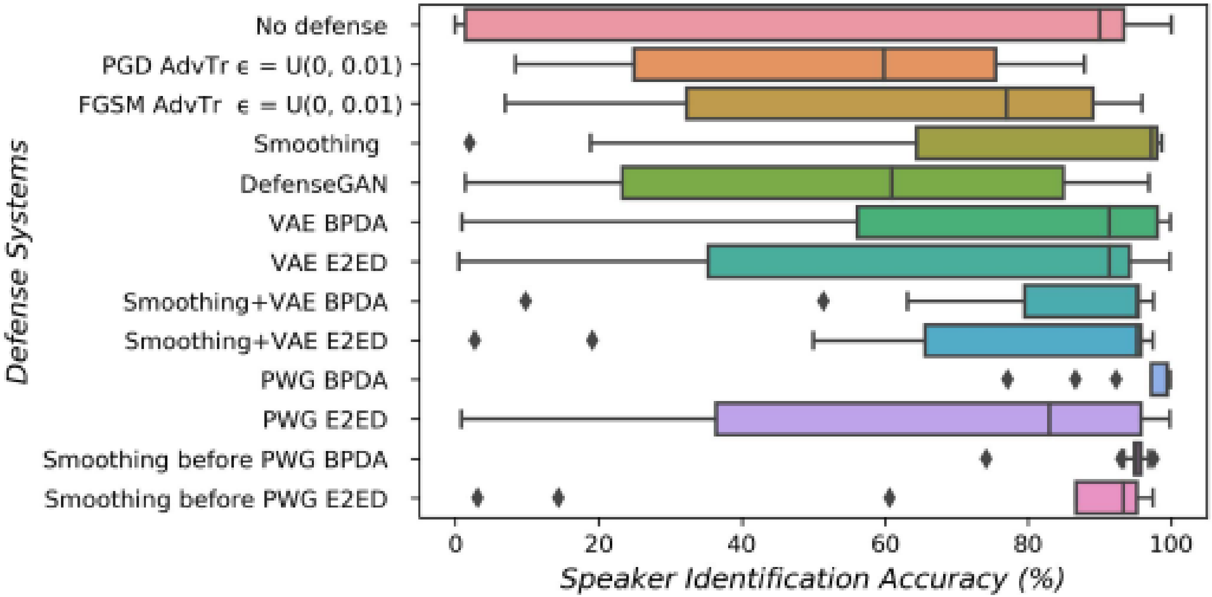


Fig. 5. Summary of all defense systems with their best settings for all attack settings as in Table VII using boxplot.

[1] Study of Pre-Processing Defenses Against Adversarial Attacks on State-of-the-Art Speaker Recognition Systems, Sonal Joshi, 2021 IEEE transactions on information forensics and security

Some Adversarial Attacks Repositories available at

https://github.com/stars/amindadgar/lists/adversarial-toolboxes

References

- Study of Pre-Processing Defenses Against Adversarial Attacks on State-of-the-Art Speaker Recognition Systems, Sonal Joshi, 2021 IEEE transactions on information forensics and security
- Ref: https://github.com/osm3000/adversarial_attack_experiment
- Universal adversarial perturbations, Seyed-Mohsen Moosavi-Dezfooli, CVPR 2017
- How to Robustify Black-Box ML Models? A Zeroth-Order Optimization Perspective, Yimeng Zhang, 2022 ICLR Spotlight

Any Questions?