TSFool: Crafting Highly-Imperceptible Adversarial Time Series through Multi-Objective Attack

Yanyun Wang [1], Dehui Du [2*], Haibo Hu [1*], Zi Liang [1] and Yuanhao Liu [2]

[1] Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong, China

[2] Software Engineering Institute, East China Normal University, Shanghai, China

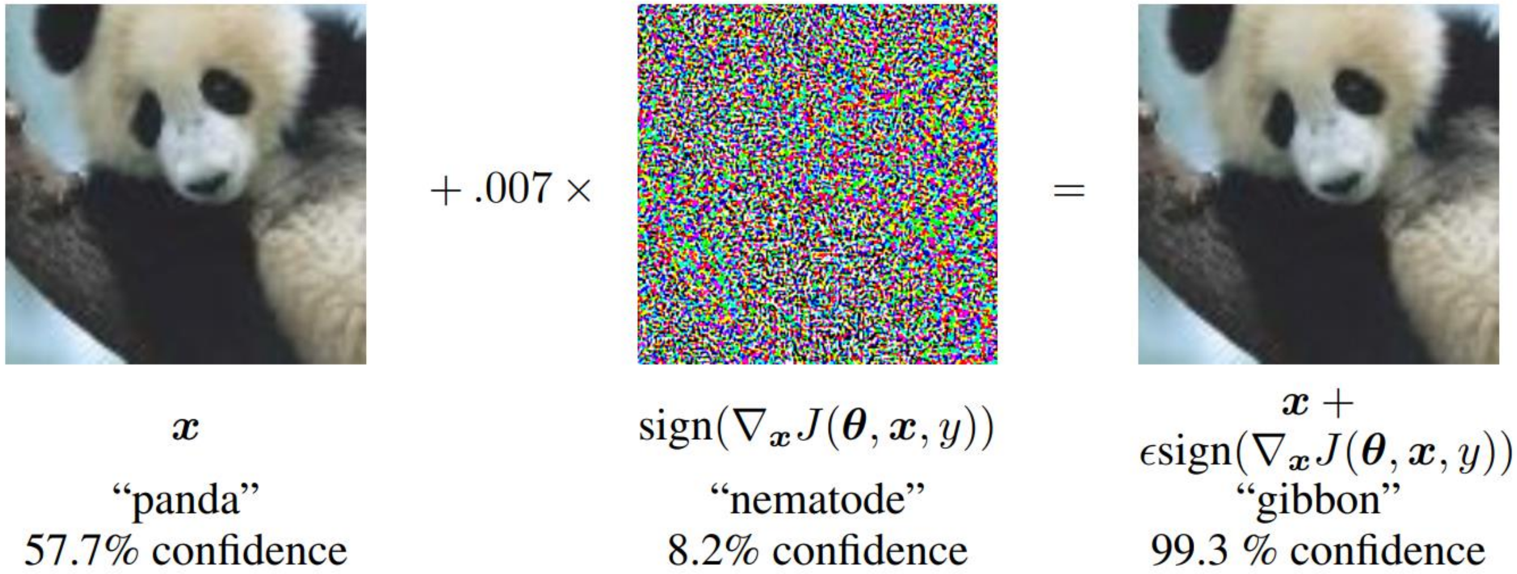
[*] Corresponding Authors

Presenter: Yanyun Wang

BACKGROUND

Topic Introduction

- Neural network (NN) classifiers are vulnerable to adversarial samples, which means imperceptible perturbations added to the input can cause the output to change significantly.



Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." In ICLR, 2015.

BACKGROUND

Existing Knowledge

- Adversarial attack is to artificially craft adversarial samples to measure the robustness of NN models.

$$\vec{x}^* = \vec{x} + \delta_{\vec{x}} = \vec{x} + \min \|\vec{z}\| \text{ s.t. } f(\vec{x} + \vec{z}) \neq f(\vec{x})$$

BACKGROUND

☐ Existing Knowledge

- Adversarial attack is to artificially craft adversarial samples to measure the robustness of NN models.

$$\vec{x}^* = \vec{x} + \delta_{\vec{x}} = \vec{x} + \min \|\vec{z}\| \text{ s.t. } f(\vec{x} + \vec{z}) \neq f(\vec{x})$$

- Gradient-based white-box adversarial attacks have achieved impressive performance on feed-forward NN classifiers and image data.

  - FGSM $\delta_{\vec{x}} = \varepsilon \, \mathrm{sign}(\nabla_{\vec{x}} \mathcal{L}(f, \vec{x}, \vec{y}))$

  - PGD $\vec{x}^{t+1} = \Pi_{\epsilon} \left\{ \vec{x}^t + \varepsilon \cdot \mathrm{sign}(\nabla_{\vec{x}} \mathcal{L}(f, \vec{x}^t, \vec{y})),\ \vec{x} \right\}$

MOTIVATION

Current Gap

- While recent years have witnessed the success of recurrent neural network (RNN) models in time series classification (TSC) tasks, the gradient-based white-box adversarial attacks cannot perform well on RNN-based TSC.

MOTIVATION

Current Gap

- While recent years have witnessed the success of recurrent neural network (RNN) models in time series classification (TSC) tasks, the gradient-based white-box adversarial attacks cannot perform well on RNN-based TSC.

Reasons

- From RNN: the unique cyclical computation in RNN architecture prevents direct model differentiation, which means the majority of gradient information is no longer directly available through the chain rule.

MOTIVATION

Current Gap

- While recent years have witnessed the success of recurrent neural network (RNN) models in time series classification (TSC) tasks, the gradient-based white-box adversarial attacks cannot perform well on RNN-based TSC.

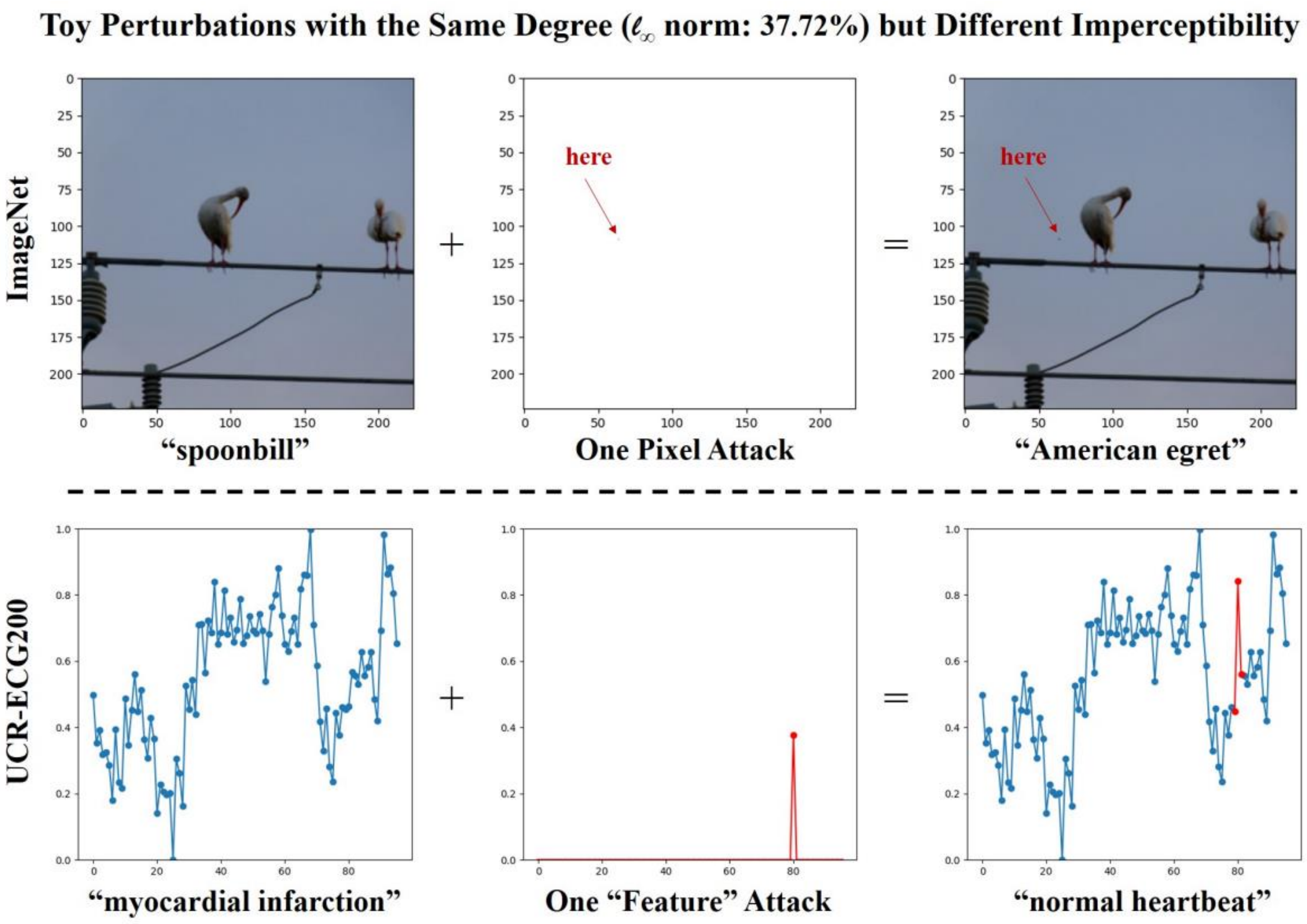
Reasons

- From RNN: the unique cyclical computation in RNN architecture prevents direct model differentiation, which means the majority of gradient information is no longer directly available through the chain rule;

- From TSC: time series data are far more visually sensitive to perturbations than image data, which poses challenges to the conventional local optimization objective of adversarial attack to minimize the perturbation amount for every single sample.

MOTIVATION

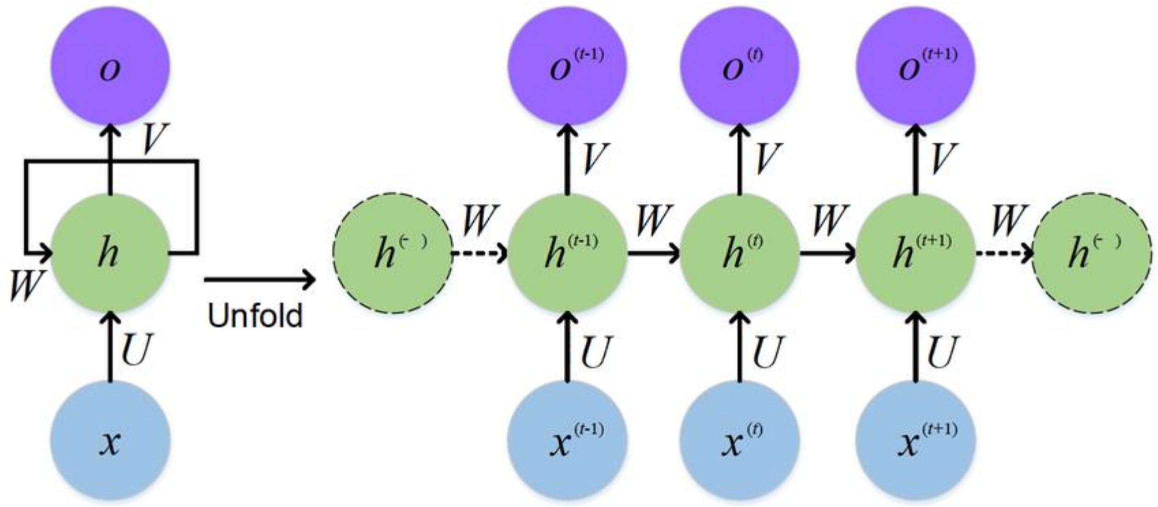
- Current Gap
  - While recent years have witne
    in time series classification (T
    cannot perform well on RNN
- Reasons
  - From RNN: the unique cyclic
    differentiation, which means
    available through the chain ru
  - From TSC: time series data a
    data, which poses challenges
    adversarial attack to minimize



MOTIVATION

Related Works

- Making RNN completely differentiable by cyclical computational graph unfolding to cater to the gradient-based methods.
  - Turns out to be inefficient and hard to stably scale in real-world practice.



MOTIVATION

Related Works

- Making RNN completely differentiable by cyclical computational graph unfolding to cater to the gradient-based methods.
  - Turns out to be inefficient and hard to stably scale in real-world practice.
- Just ignoring the given model knowledge and implementing non-gradient-based black-box attacks.

Related Works

- Making RNN completely differentiable by cyclical computational graph unfolding to cater to the gradient-based methods.
  - Turns out to be inefficient and hard to stably scale in real-world practice.
- Just ignoring the given model knowledge and implementing non-gradient-based black-box attacks.
  - By model querying:
    - A majority of these methods like One Pixel Attack and Square Attack can only work for image input.

MOTIVATION

Related Works

- Making RNN completely differentiable by cyclical computational graph unfolding to cater to the gradient-based methods.
    - Turns out to be inefficient and hard to stably scale in real-world practice.
- Just ignoring the given model knowledge and implementing non-gradient-based black-box attacks.
    - By model querying:
        - A majority of these methods like One Pixel Attack and Square Attack can only work for image input; and
        - The rest of these methods like Boundary Attack and HopSkipJump are extremely time-consuming because they rely on random-walking.

MOTIVATION

Related Works

- Making RNN completely differentiable by cyclical computational graph unfolding to cater to the gradient-based methods.
  - Turns out to be inefficient and hard to stably scale in real-world practice.
- Just ignoring the given model knowledge and implementing non-gradient-based black-box attacks.
  - By model querying:
    - A majority of these methods like One Pixel Attack and Square Attack can only work for image input; and
    - The rest of these methods like Boundary Attack and HopSkipJump are extremely time-consuming because they rely on random-walking.
  - By adversarial transferability:
    - Tends to achieve reasonable time and small perturbation, but the worst attack success rate.

MOTIVATION

- Target

MOTIVATION

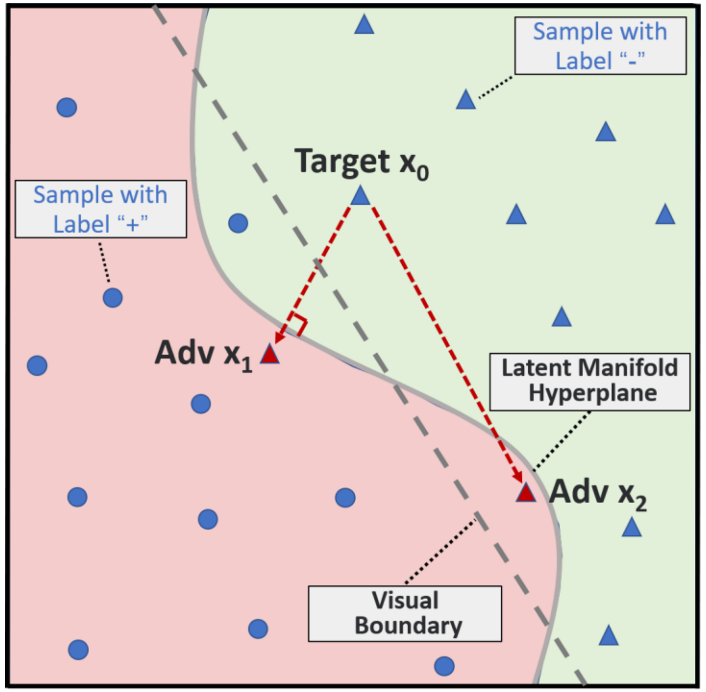
- **Target**
  - A non-gradient-based adversarial attack method (due to RNN model).

MOTIVATION

- **Target**
  - A non-gradient-based adversarial attack method (due to RNN model),
  - With additional consideration for the imperceptibility of perturbation (due to time series data).

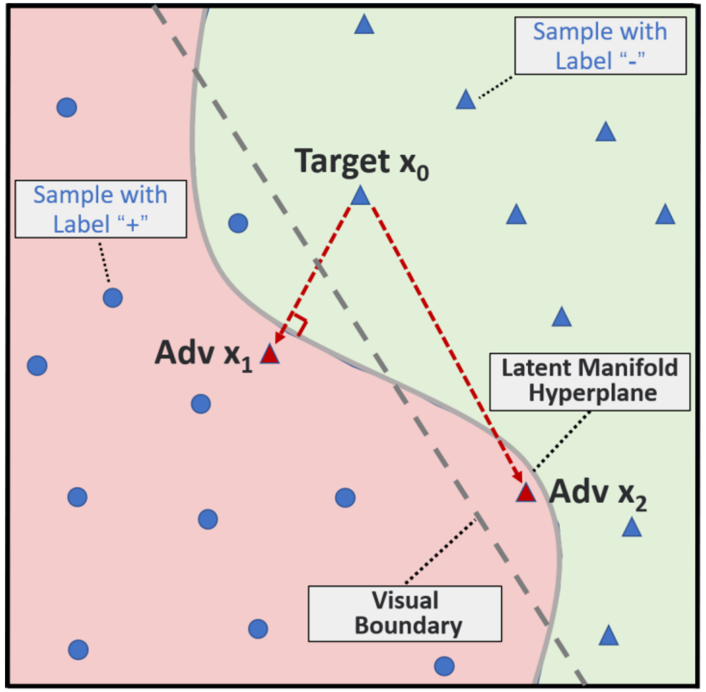
CORE IDEA

- Explanation for adversarial sample based on Manifold Hypothesis



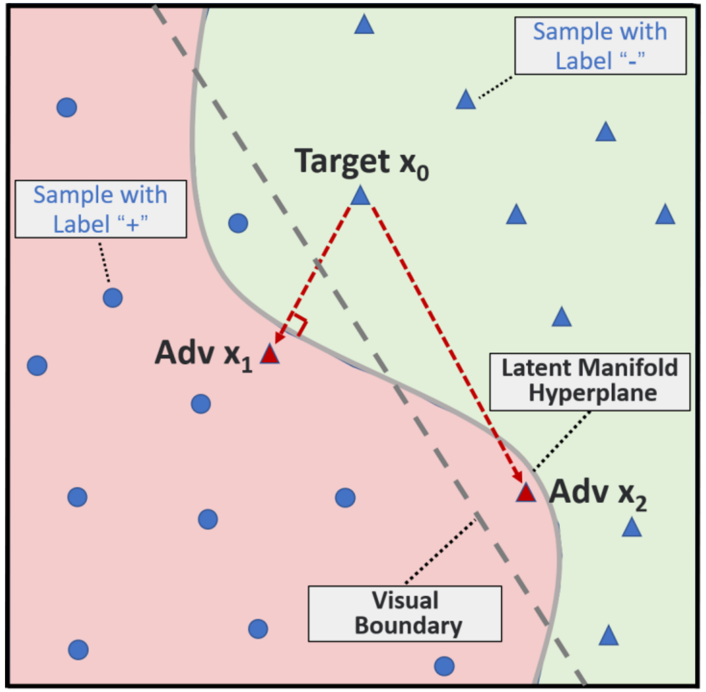
CORE IDEA

- **Explanation for adversarial sample based on Manifold Hypothesis**
  - NN-based classification relies on the latent manifold hyperplane (i.e., the classification boundary).



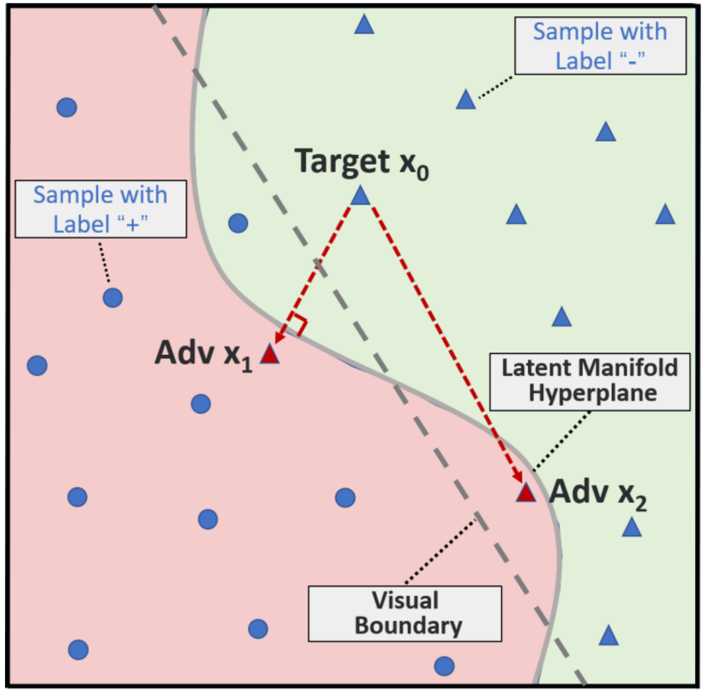
CORE IDEA

- **Explanation for adversarial sample based on Manifold Hypothesis**
  - NN-based classification relies on the latent manifold hyperplane (i.e., the classification boundary);
  - The features of input data cannot always semantically and visually reflect the latent manifold.



CORE IDEA

- **Explanation for adversarial sample based on Manifold Hypothesis**
  - NN-based classification relies on the latent manifold hyperplane (i.e., the classification boundary);
  - The features of input data cannot always semantically and visually reflect the latent manifold;
  - A small perturbation in human cognition imposed on a sample may completely overturn the perception of NN to its latent manifold.



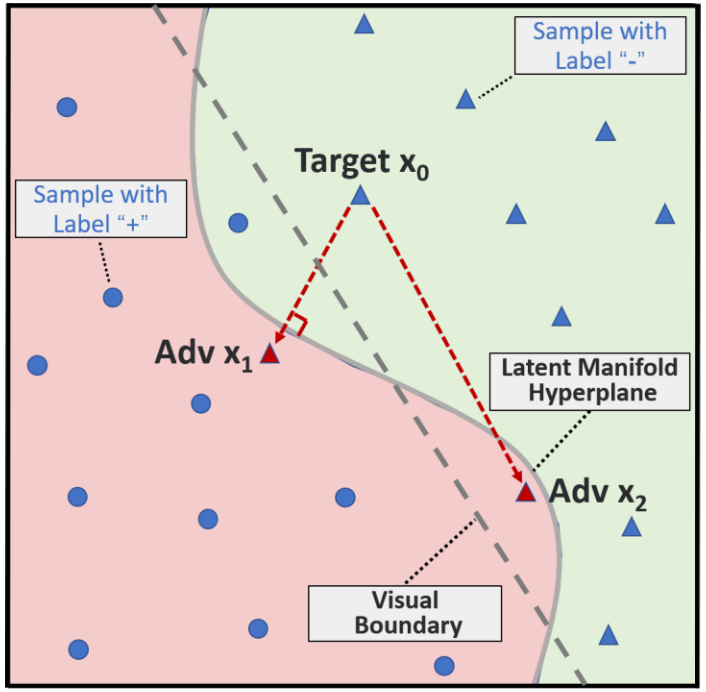
CORE IDEA

☐ Explanation for adversarial sample based on Manifold Hypothesis

- NN-based classification relies on the latent manifold hyperplane (i.e., the classification boundary);
- The features of input data cannot always semantically and visually reflect the latent manifold;
- A small perturbation in human cognition imposed on a sample may completely overturn the perception of NN to its latent manifold.

☐ Arguments

- Even the minimal local perturbation is not necessarily the most imperceptible one from the global perspective.



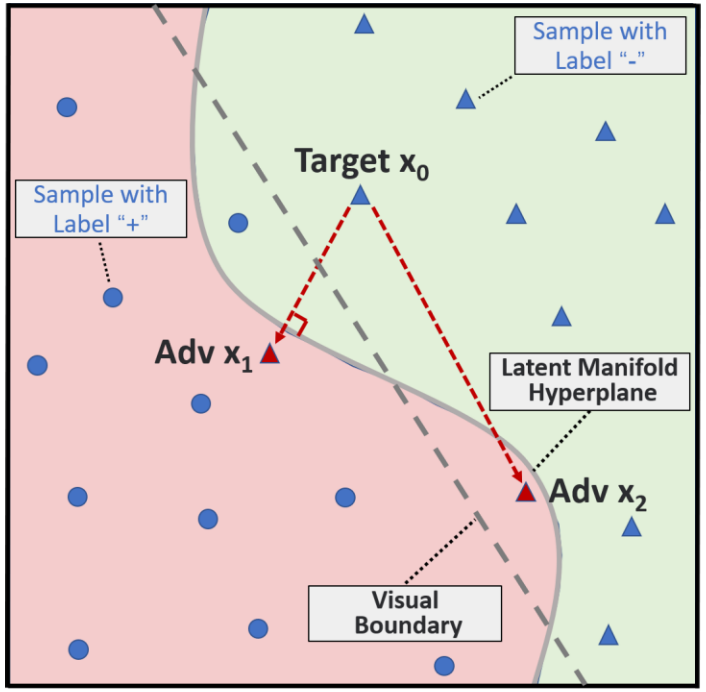
CORE IDEA

Explanation for adversarial sample based on Manifold Hypothesis

- NN-based classification relies on the latent manifold hyperplane (i.e., the classification boundary);
- The features of input data cannot always semantically and visually reflect the latent manifold;
- A small perturbation in human cognition imposed on a sample may completely overturn the perception of NN to its latent manifold.

Arguments

- Even the minimal local perturbation is not necessarily the most imperceptible one from the global perspective;
- The conventional approach to approximate the local optimization objective does not always lead to a highly-imperceptible adversarial attack.



CORE IDEA

Camouflage Coefficient

- A novel global optimization objective that takes the relative position between adversarial samples and class clusters into consideration, to measure the imperceptibility of adversarial samples from the perspective of class distribution.

CORE IDEA

Camouflage Coefficient

- A novel global optimization objective that takes the relative position between adversarial samples and class clusters into consideration, to measure the imperceptibility of adversarial samples from the perspective of class distribution.

$$\mathcal{C}(\vec{x}^*) = \frac{\|\vec{x}^* - \vec{m}_i\|/d_i}{\|\vec{x}^* - \vec{m}_j\|/d_j}$$

$$\vec{m}_i = \frac{1}{|\mathcal{X}_i|} \sum_{\vec{x}' \in \mathcal{X}_i} \vec{x}'$$

$$d_i = \frac{1}{|\mathcal{X}_i|} \sum_{\vec{x}' \in \mathcal{X}_i} \|\vec{x}' - \vec{m}_i\|$$

- It is the relative proportion of the norm distance between the adversarial sample and the original class to the distance between it and the misclassified class regarding the different cluster ranges of the classes.

CORE IDEA

Multi-objective Optimization Problem

- By adding the Camouflage Coefficient, we refine the adversarial attack task to a multi-objective optimization problem.

CORE IDEA

Multi-objective Optimization Problem

- By adding the Camouflage Coefficient, we refine the adversarial attack task to a multi-objective optimization problem.

Approximative Solution

CORE IDEA

- **Multi-objective Optimization Problem**
  - By adding the Camouflage Coefficient, we refine the adversarial attack task to a multi-objective optimization problem.
- **Approximative Solution**
  - To cross the manifold hyperplane in the position that is:
    - The closest to the benign sample.

CORE IDEA

Multi-objective Optimization Problem

- By adding the Camouflage Coefficient, we refine the adversarial attack task to a multi-objective optimization problem.

Approximative Solution

- To cross the manifold hyperplane in the position that is:
  - Sufficiently close to the benign sample.

CORE IDEA

Multi-objective Optimization Problem

- By adding the Camouflage Coefficient, we refine the adversarial attack task to a multi-objective optimization problem.

Approximative Solution

- To cross the manifold hyperplane in the position that is:
  - Sufficiently close to the benign sample; and also
  - Sufficiently close to the center of mass of the benign class (i.e., m_i).

CORE IDEA

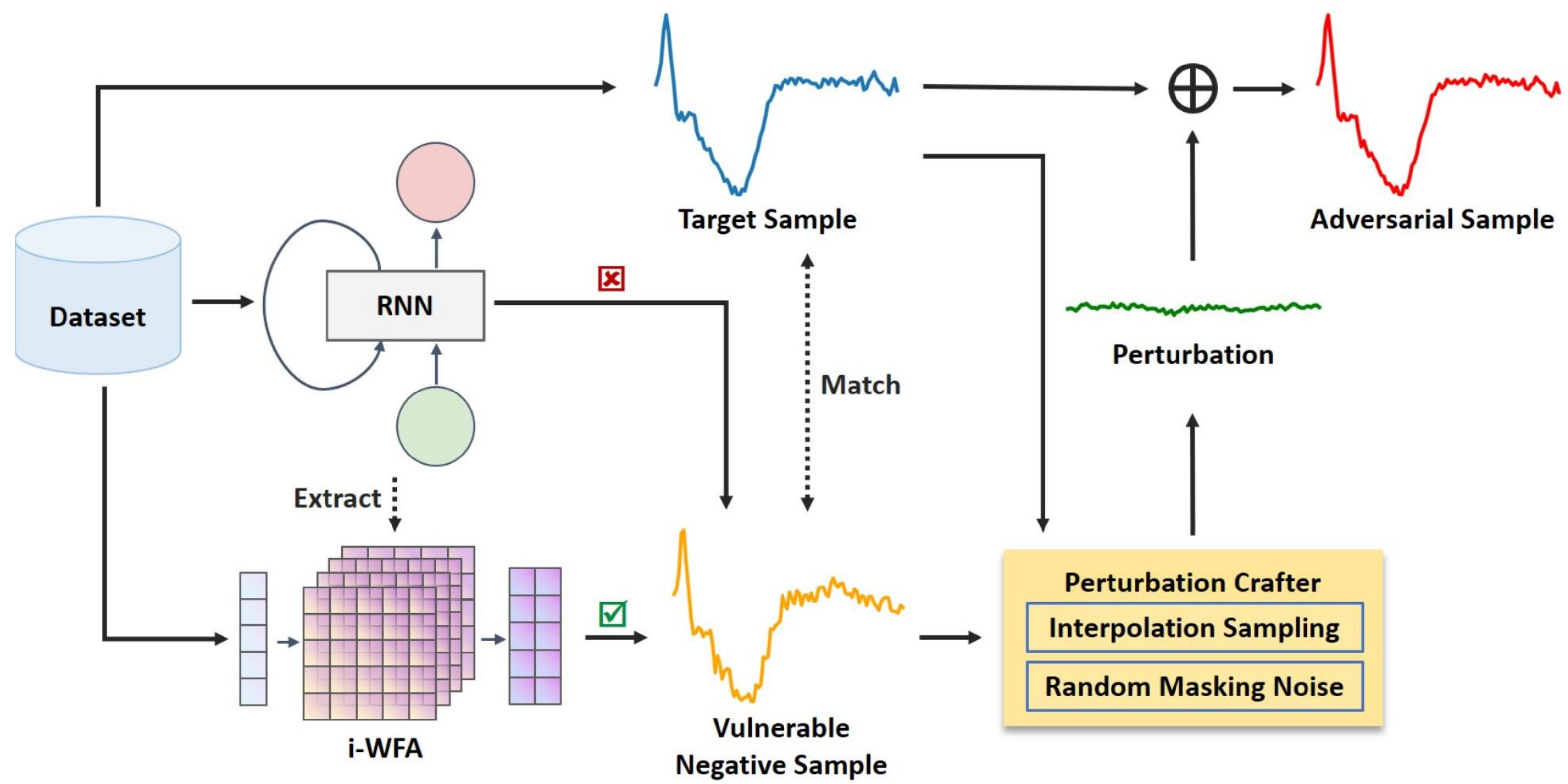
Multi-objective Optimization Problem

- By adding the Camouflage Coefficient, we refine the adversarial attack task to a multi-objective optimization problem.

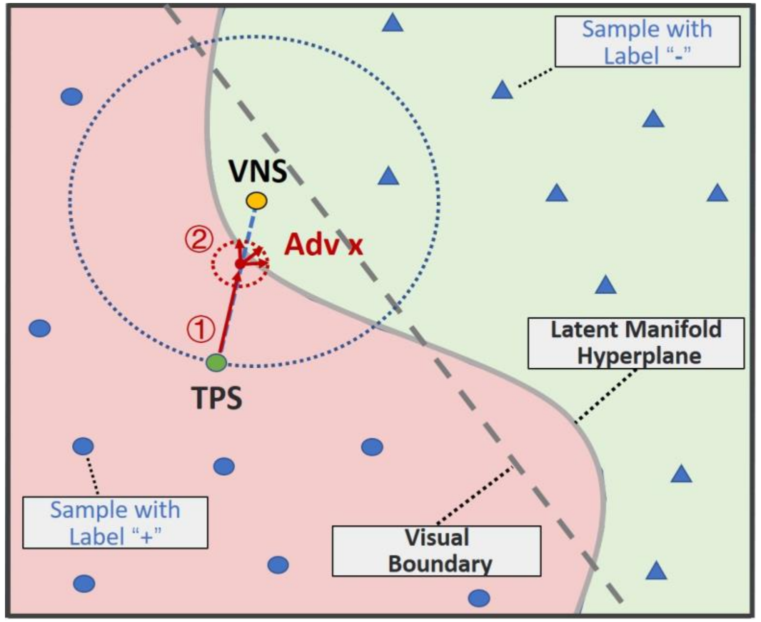
Approximative Solution

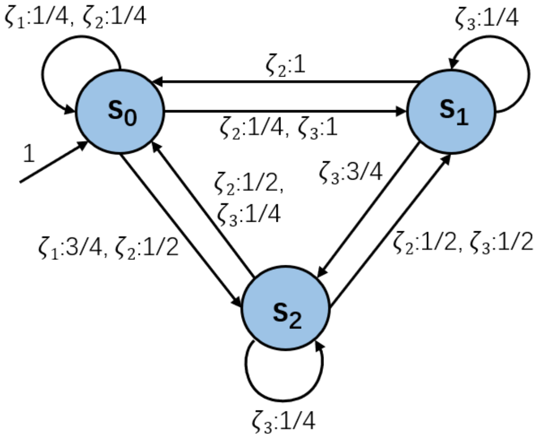
- To cross the manifold hyperplane in the position that is:
  - Sufficiently close to the benign sample; and also
  - Sufficiently close to the center of mass of the benign class (i.e., m_i),
- In a non-gradient-based way.

METHOD



$$\delta_{\vec{x}} = \lambda_\varepsilon \|\vec{x}_s - \vec{x}\| + \vec{x}_\varepsilon$$



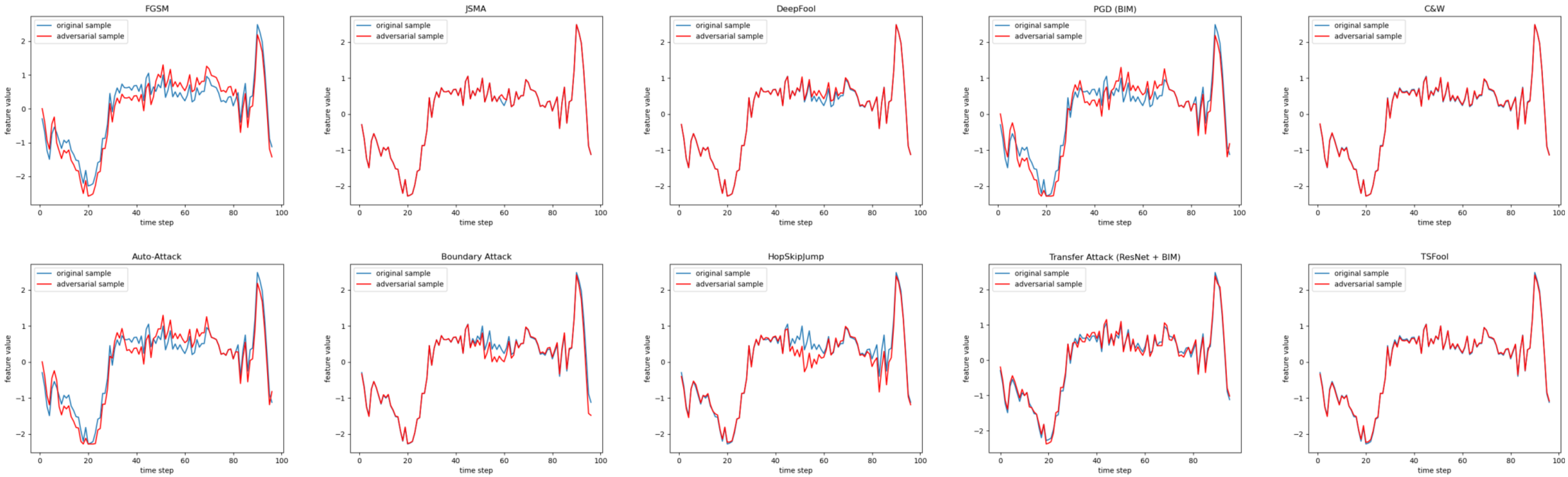


The representation model named i-WFA is built only upon the RNN's outputs. It can fit the manifold hyperplane of an RNN classifier but distinguish samples by their original features like humans. As a result, it can capture deeply embedded vulnerable samples whose features deviate from the latent manifold.

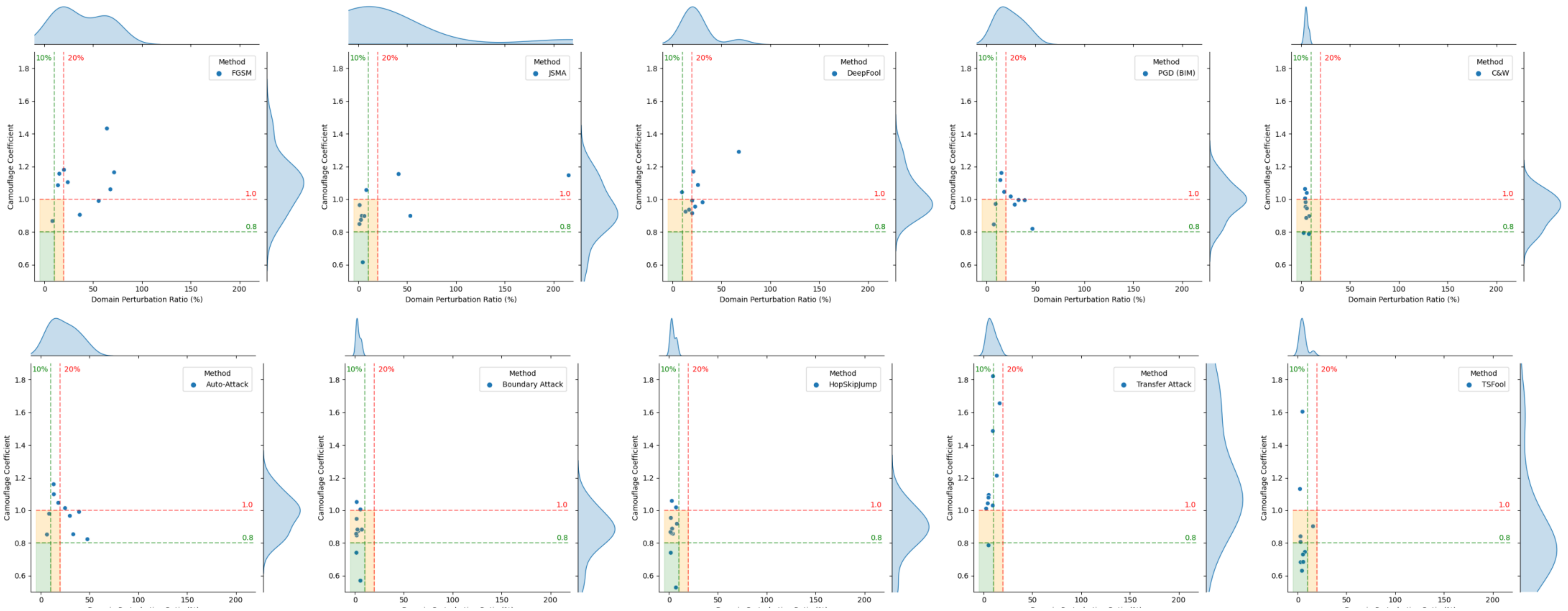
EVALUATION

Method	Attack Success Rate	Generation Number	Average Time Cost (s)	Perturbation Ratio (ρ^*)	Camouflage Coefficient
FGSM	72.12%	300.75	0.0018	37.13%	1.0804
JSMA	83.53%		1.0287	15.06%	0.9476
DeepFool	81.58%		0.0276	21.45%	1.0107
PGD (BIM)	76.84%		0.1327	22.71%	0.9938
C&W	69.90%		3.2016	5.16%	0.9372
Auto-Attack	80.11%		0.1824	22.55%	0.9745
Boundary Attack	79.01%		9.0399	3.04%	0.8788
HopSkipJump	83.17%		12.3068	3.86%	0.8872
Transfer Attack	19.54%	250	-	7.68%	1.2010
TSFool	**87.76%**	305	**0.0230**	**4.63%**	**0.8147**

EVALUATION

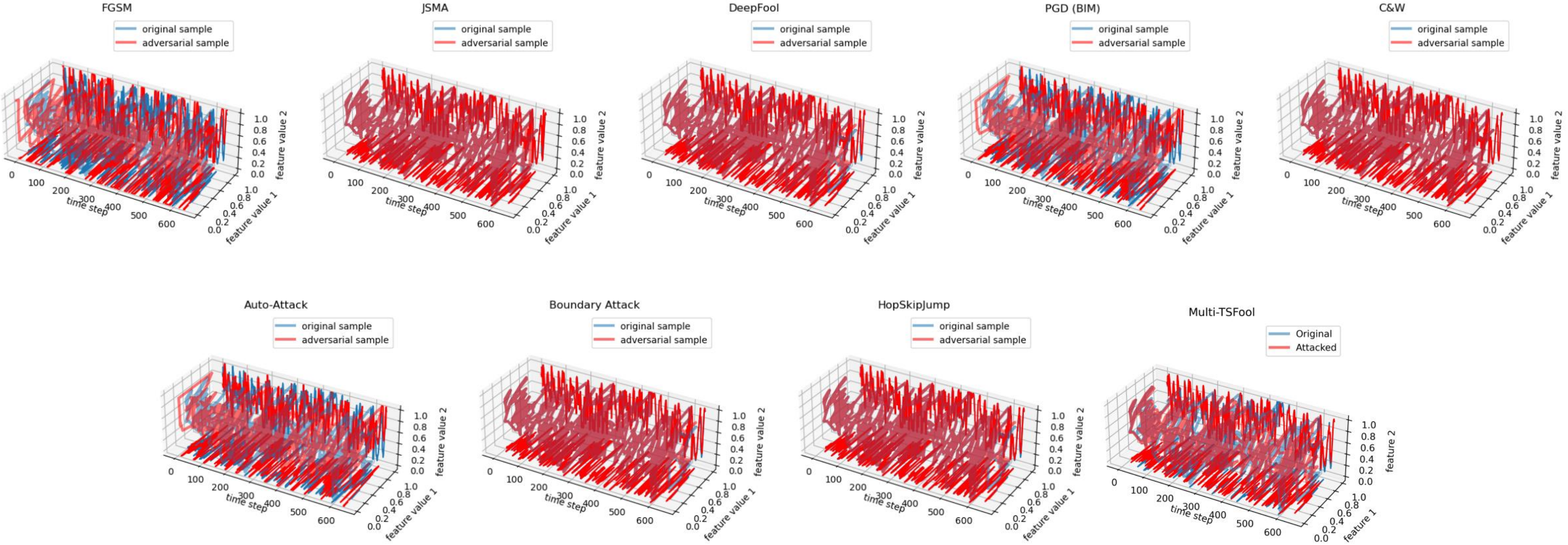

FGSM
JSMA
DeepFool
PGD (BIM)
C&W
Auto-Attack
Boundary Attack
HopSkipJump
Transfer Attack (ResNet + BIM)
TSFool
original sample
adversarial sample
feature value
time step

EVALUATION

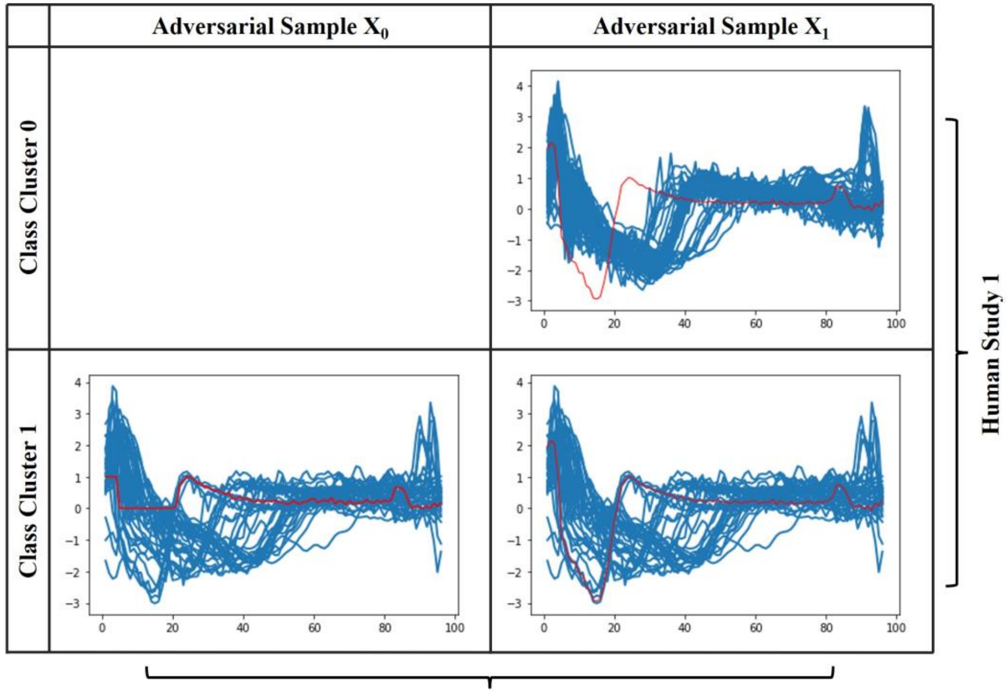

Method
FGSM
JSMA
DeepFool
PGD (BIM)
C&W
Auto-Attack
Boundary Attack
HopSkipJump
Transfer Attack
TSFool
10%
20%
1.0
0.8
Camouflage Coefficient
Domain Perturbation Ratio (%)

EVALUATION

Target Model		Method	Attack Success Rate	Generation Number	Average Time Cost (s)	Perturbation Ratio (ρ^*)			Camouflage Coefficient
Dataset	Accuracy					ℓ_1	ℓ_2	ℓ_∞	
AF	0.8000	FGSM	80.00%	15	0.0127	24.52%	25.87%	29.04%	1.0211
		JSMA	86.67%		7.4301	3.04%	3.74%	4.99%	0.8126
		DeepFool	80.00%		0.9900	0.54%	0.64%	0.85%	0.7551
		PGD (BIM)	80.00%		0.9406	16.79%	18.22%	21.72%	0.9918
		C&W	40.00%		13.2858	0.48%	0.52%	0.63%	0.7958
		Auto-Attack	80.00%		2.1487	15.80%	17.11%	20.35%	0.9918
		Boundary Attack	66.67%		418.2511	0.43%	0.51%	0.67%	0.8045
		HopSkipJump	86.67%		78.3258	0.84%	0.97%	1.27%	0.8066
		TSFool	100.00%	20	0.0960	5.89%	6.69%	8.65%	0.6047

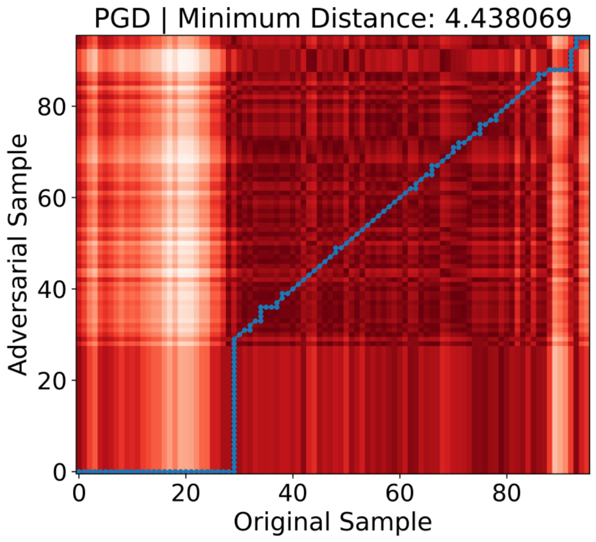


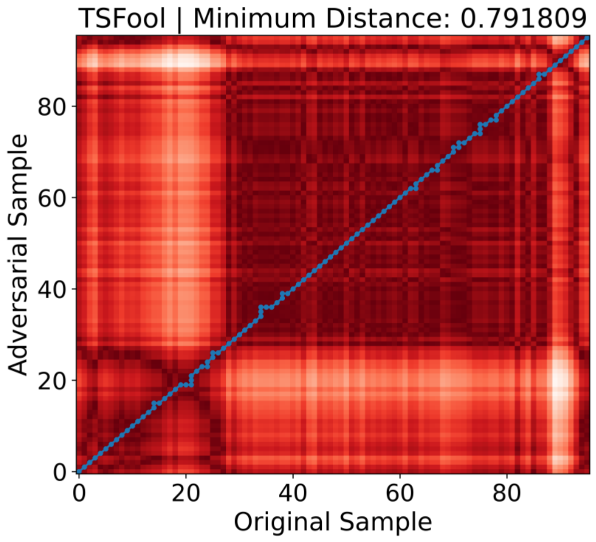
EVALUATION



Study	Option	Question							Sum	Count
		1	2	3	4	5	6	7		
Study 1	The Original Class Cluster	20	13	55	37	47	55	31	258	5
	The Misclassified Class Cluster	41	48	4	22	13	8	28	164	2
	Neutral	4	4	6	6	5	2	6	33	0
Study 2	The Adversarial Sample from TSFool	54	58	51	57	58	60	58	396	7
	The Adversarial Sample from PGD	10	5	12	6	4	3	5	45	0
	Neutral	1	2	2	2	3	2	2	14	0

EVALUATION





Method	Metric	Time Series Anomaly Detection			
		OCSVM	IF	LOF	LSTMOD
PGD (BIM)	Pre	0.1755	0.1899	0.2794	0.2107
	Re	0.4890	0.5377	0.7018	0.7091
	$F1$	0.2454	0.2692	0.3782	0.3092
C&W	Pre	0.0637	0.0534	0.0463	0.0968
	Re	0.1304	0.1201	**0.0745**	0.3377
	$F1$	0.0798	0.0696	**0.0534**	0.1432
HopSkipJump	Pre	0.0693	0.0473	0.0801	0.1381
	Re	0.1561	0.1115	0.2282	0.5127
	$F1$	0.0897	0.0640	0.1146	0.2129
TSFool	Pre	**0.0505**	**0.0346**	**0.0460**	**0.0741**
	Re	**0.1012**	**0.0829**	0.1274	**0.3218**
	$F1$	**0.0622**	**0.0469**	0.0639	**0.1175**

CONCLUSION

- **Insights**

- **Contributions**

CONCLUSION

Insights

- General consideration beyond image data and feed-forward models is still lacking in the current knowledge of adversarial attack.

Contributions

CONCLUSION

- **Insights**
  - General consideration beyond image data and feed-forward models is still lacking in the current knowledge of adversarial attack;
  - Imperceptibility measures of adversarial samples have not received sufficient attention, without which it would be hard to fairly define “adversarial”.

- **Contributions**

CONCLUSION

Insights

- General consideration beyond image data and feed-forward models is still lacking in the current knowledge of adversarial attack;
- Imperceptibility measures of adversarial samples have not received sufficient attention, without which it would be hard to fairly define “adversarial”.

Contributions

- A novel global optimization objective "Camouflage Coefficient" to refine the adversarial attack as a multi-objective optimization problem.

CONCLUSION

Insights

- General consideration beyond image data and feed-forward models is still lacking in the current knowledge of adversarial attack;
- Imperceptibility measures of adversarial samples have not received sufficient attention, without which it would be hard to fairly define “adversarial”.

Contributions

- A novel global optimization objective "Camouflage Coefficient" to refine the adversarial attack as a multi-objective optimization problem;
- A new latent manifold-based methodology to heuristically approximate the solution of the suggested optimization problem, which opens a new feasible path to craft imperceptible adversarial samples.

Thank You!

Acknowledgements

This work was supported by National Natural Science Foundation of China (Grant No: 92270123, 62072390), and the Research Grants Council, Hong Kong SAR, China (Grant No: PolyU 15203120, 15226221, 15209922, and 15210023).

Contact Information

Haibo Hu: haibo.hu@polyu.edu.hk & Dehui Du: dhdu@sei.ecnu.edu.cn

The Hong Kong Polytechnic University 香港理工大學

華東師範大學 East China Normal University