CommSec

HITB SECCONF AMSTERDAM - 2021

Model Robustness Will Hurt Data Privacy?

Jiqiang Gao, Mengyun Tang, Tony Huang

Tencent Zhuque Lab

Who We Are

- From Tencent Zhuque Lab of Security Platform Dpt.
- **Tencent Security Platform Dpt**. Has been with Tencent for 16 years, and dedicated to the protection of QQ, Wechat, Tencent Games and other critical products.
- Focus on Tencent accounts security, AI security, anti-fraud, anti-scalping, intrusion detection, and mobile app security, etc.
- **Tencent Zhuque Lab** was founded in 2019 by Tencent Security Platform Dpt., focusing on red teaming and **AI security** research.

Tencent Security Platform Dpt.

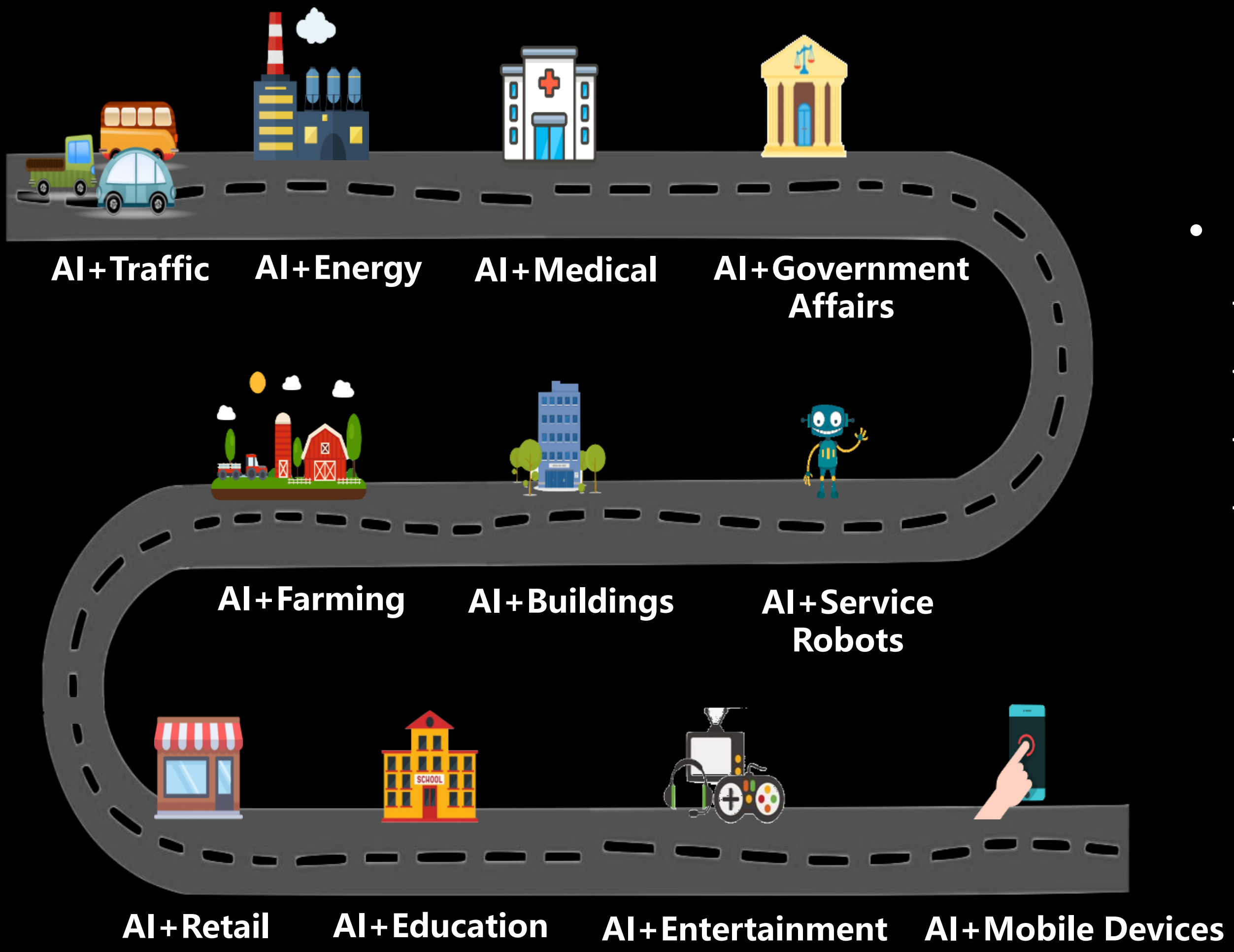
Tencent Zhuque Lab

Outline

1. AI and Security
2. Background and Motivation
   - Adversarial Attacks and Adversarial Training
   - Model Privacy Attacks
3. How to Steal Data from Model Gradient?
4. Discussion
5. Conclusion
6. Appendix： Other Interesting Study

1. AI and Security

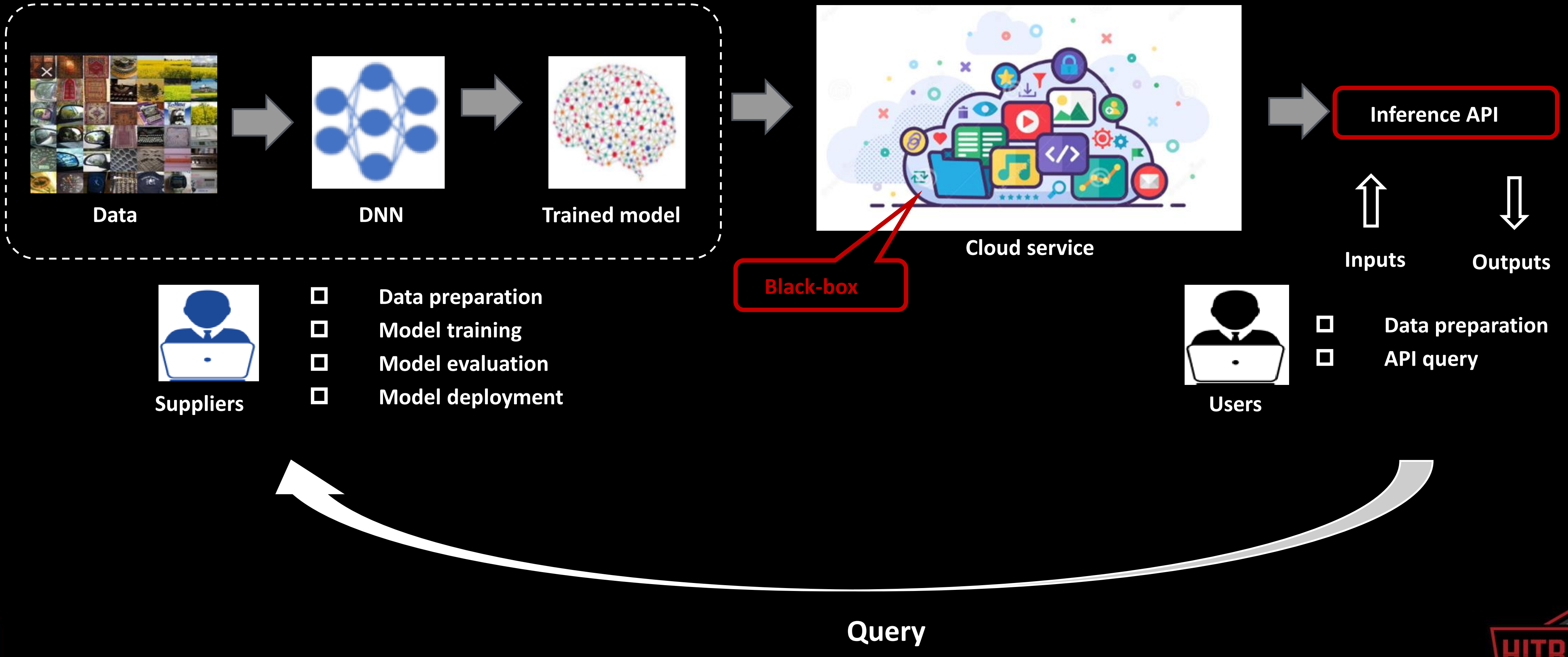
✓ **Success of AI**



- **AI is becoming a general tool**
  - No domain knowledge required
  - Can handle big data
  - Improve performance
  - Scalability

1. AI and Security

✓ Working flow

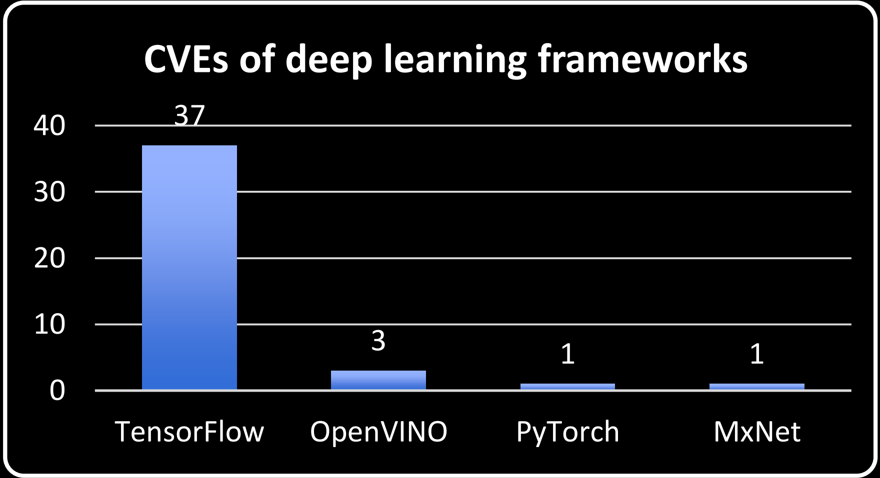


1. AI and Security

✓ Security challenges AI

- **Vulnerabilities in AI components**

  - Deep learning frameworks: TensorFlow, Caffe, MxNet, PyTorch, etc.
  - Acceleration frameworks: TensorRT, etc.
  - Software packages: OpenCV, Numpy, Pandas, etc.
  - Computing power: GPU, CPU, FPGA



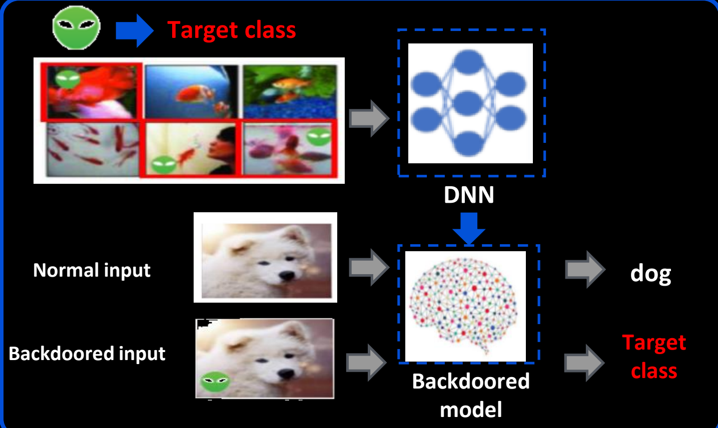
(Ref: https://cve.mitre.org)

1. AI and Security

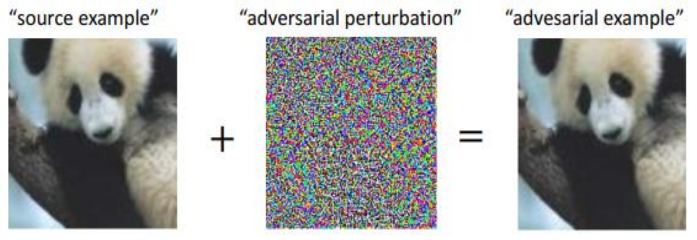
✓ **Security challenges AI**

- **New attacks targeting AI systems**
  - Data poisoning attacks
  - Backdoor attack
  - Model stealing attacks
  - …

Data | **Data poisoning attack**



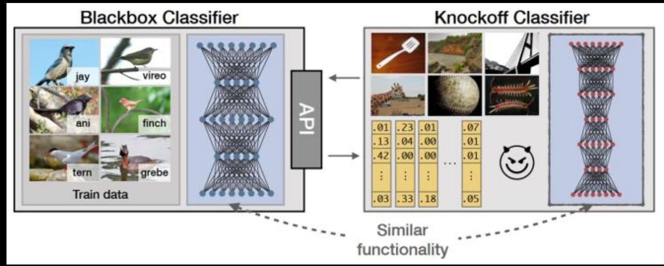
Algorithm | **Adversarial attack**



- Tiny perturbations on input
- Large perturbations on output

(Ref: Goodfellow et al, 2014)

Model | **Model stealing attack**



- Data in, prediction out
- Build a shadow model

(Ref: Tribhuvanesh Orekondy et al, 2019)

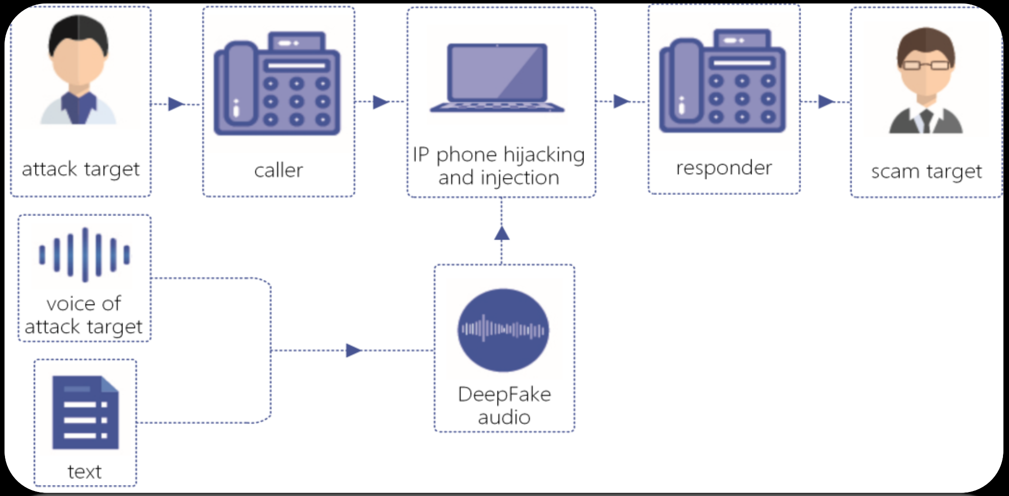
1. AI and Security

✓ Security challenges AI

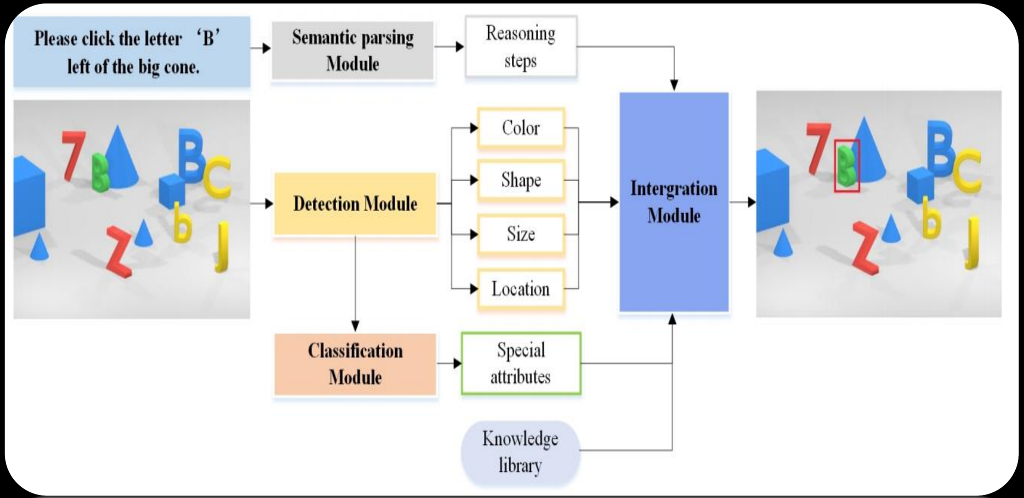
- **The abuse of AI technology**

  - Deepfake attacks

  - CAPTCHA recognition



(Ref: Mengyun Tang et al, CanSecWest 2021)



(Ref: YiPeng Gao et al, 2021)

1. AI and Security

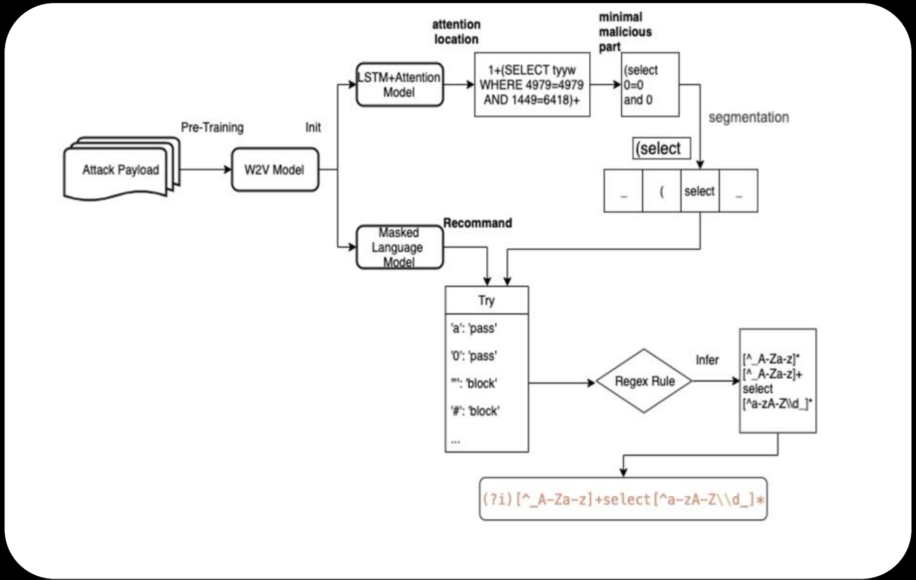
✓ AI enables security

- **Steal WAF Protection Rules**

  - Manual Methods
    - Based on expert experience
    - Observe WAF response by sending attack payload
    - Infer the rules through multiple attempts

  - Using AI
    - Use the pre-trained method to learn the security experience contained in the payload
    - Use the attention mechanism to locate the part of the payload that contributes to the detection result
    - Use the recommendation model to rank the probabilities of the candidate characters

  - Effectiveness
    - Without excessive manual intervention
    - Batch and large-scale execution



(Ref: Keyun Luo et al, Freebuf CIS 2020)

2. Background and Motivation

✓ Adversarial attacks

- **Adversarial examples**

  - Tiny perturbation on input, large perturbation on prediction
  - Easy to generate such perturbation, e.g. Fast Gradient Sign Method (FGSM) $x^* = x + sign(\nabla_x J(x, y))$
  - Exist in various AI tasks, such as image classification, object detection, and ASR, etc
  - Reveals the vulnerabilities of AI models based on deep neural networks

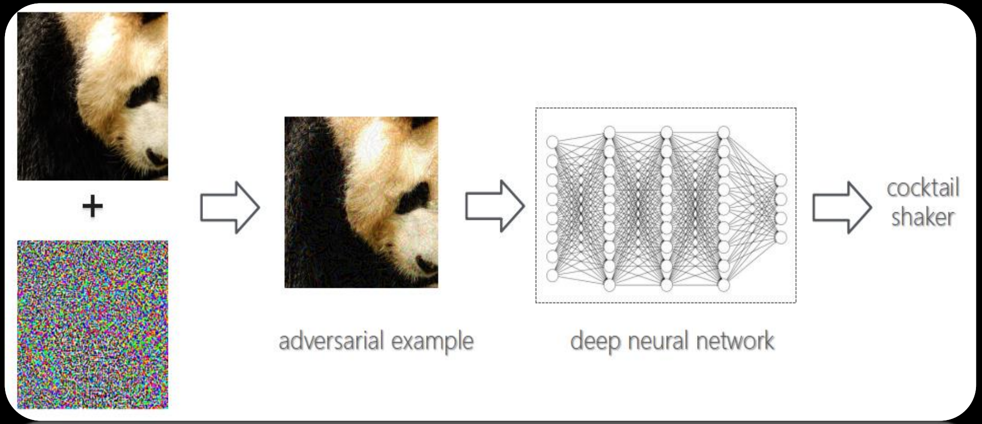
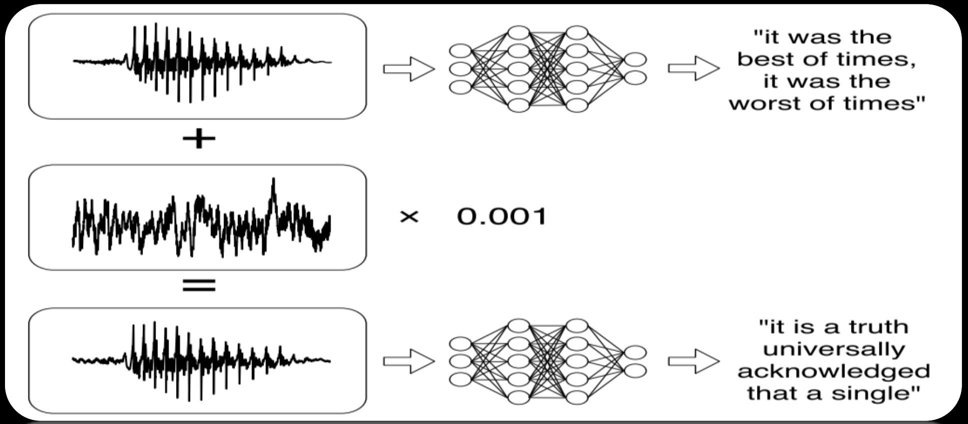


Image adversarial attack
(Ref: Mengyun Tang et al, CanSecWest 2019)



Audio adversarial attack
(Ref: Nicholas Carlini et al, S&P Workshop 2018)

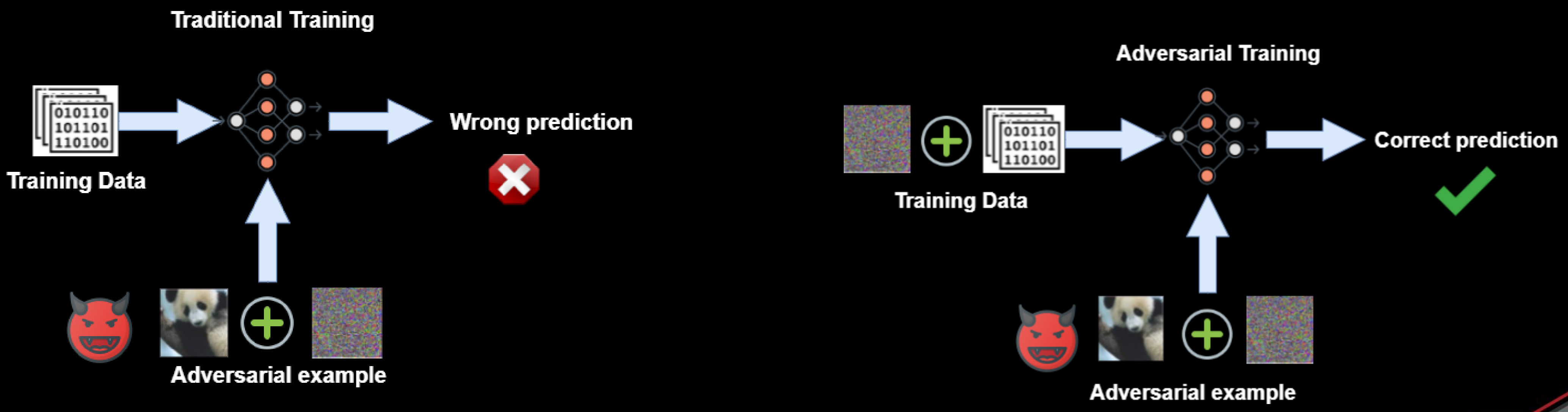
2. Background and Motivation

✓ Adversarial training

- **Adversarial training**

  - Adversarial examples are produced as a part of training data
  - Can be formulated as solving a min-max optimization problem
  - Adversarial training is one of the most promising ways to defend against adversarial attacks

$$min \frac{1}{N}\sum_{i=1}^{N} \max_{\|x_i' - x_i\|_p \le \varepsilon} l(h_\theta(x_i'), y_i)$$

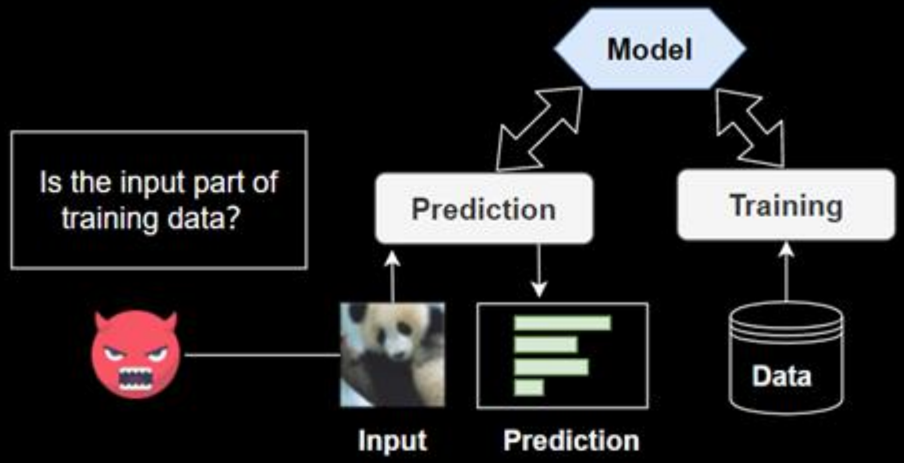


(Ref: Aleksander Madry et al. 2017)

2. Background and Motivation

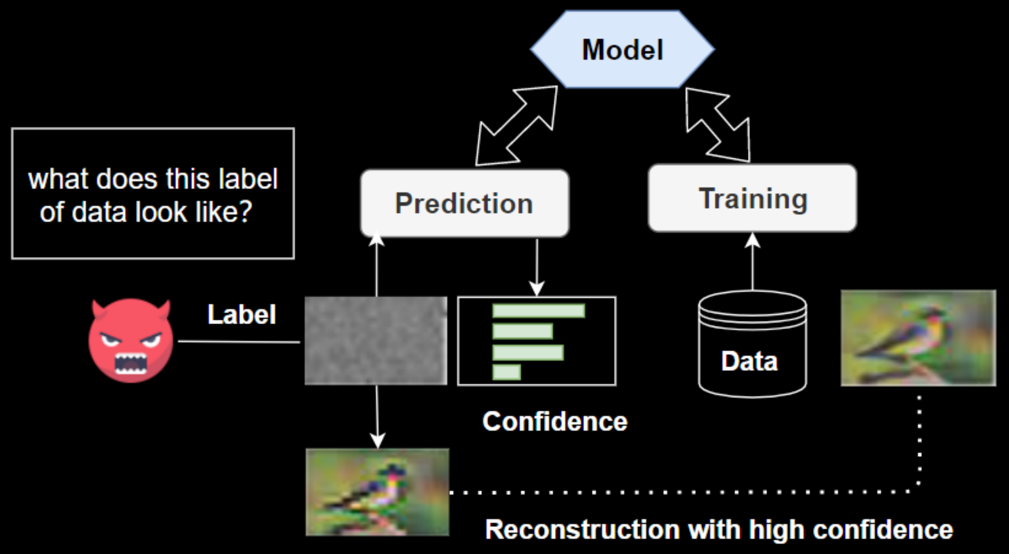
✓ Model Privacy Attacks

Membership Inference Attack



(Ref: Reza Shokri et al. 2017)

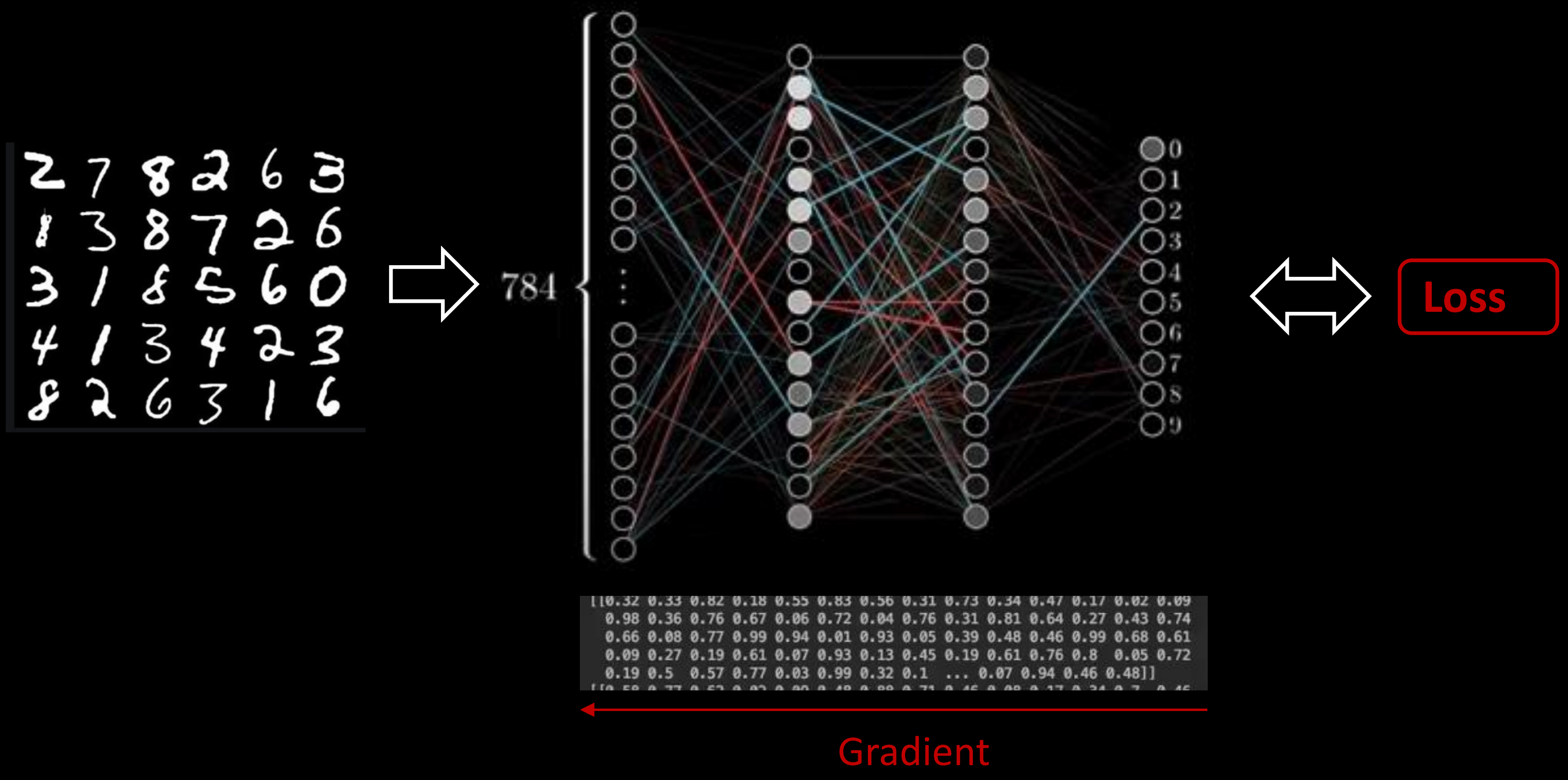
Model Inversion Attack



(Ref: Matt Fredrikson et al. 2015)

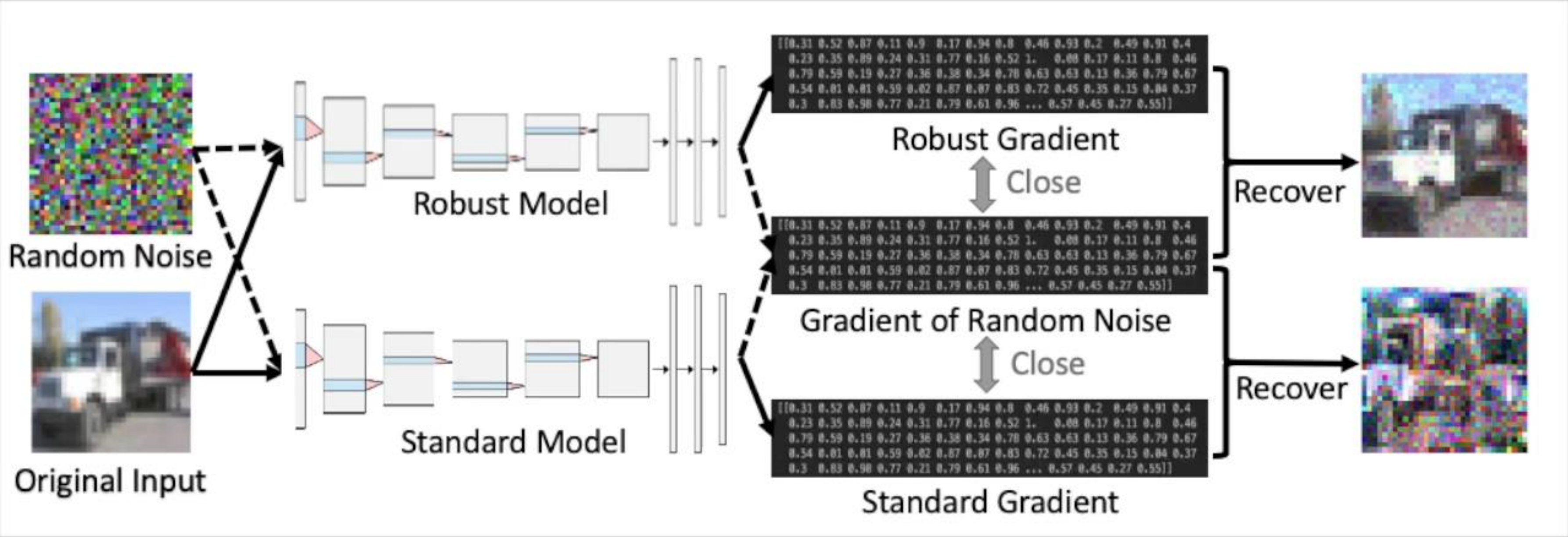
3. How to Steal Data from Model Gradient?

- ✓ **Model gradient**



3. How to Steal Data from Model Gradient?

✓ **Methodology**



4. Experiments and Discussion

✓ Preparation

- Settings

Dataset	Model	Norm	Epsilon	Step size	Iterations
CIFAR	PreActResNet18 ResNet18 WideResNet34-10	PGD_inf	0.0314	0.0078	10
MNIST	SmallCNN	PGD_inf	0.3	0.01	40

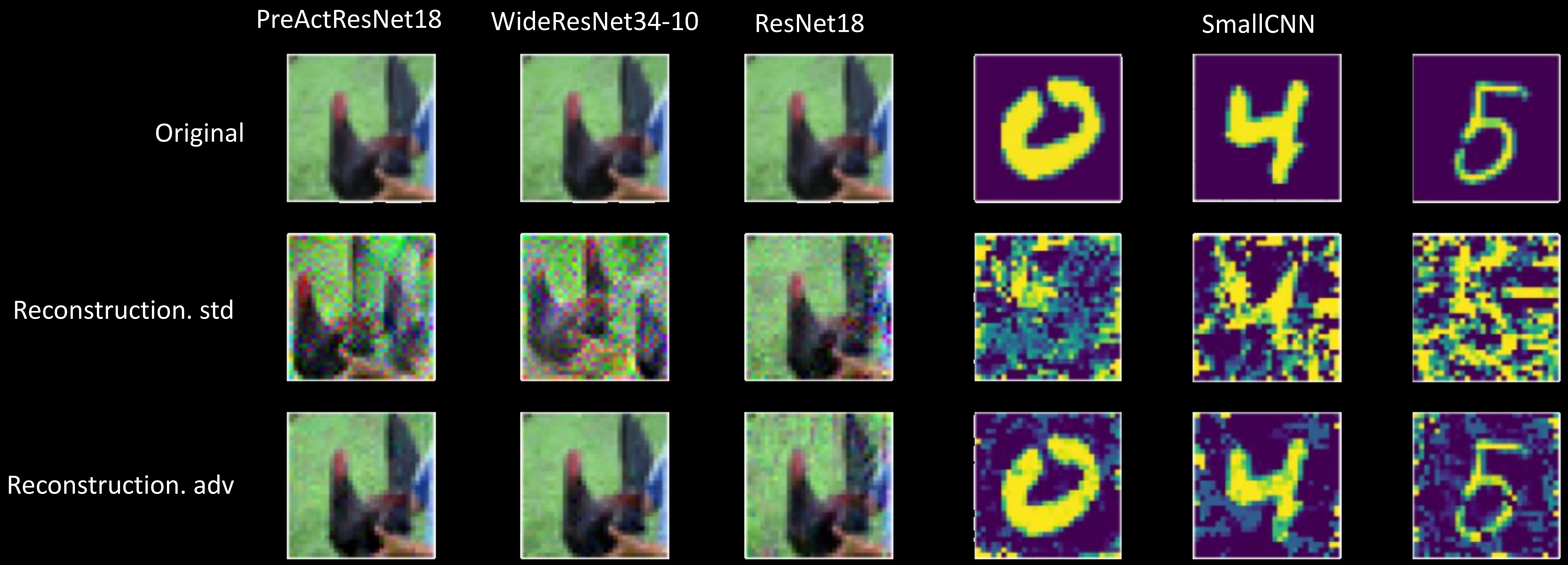
- Evaluation metrics

  - Peak Signal-to-Noise Ratio (PSNR)
  - Mean Squared Error (MSE)
  - Feature Mean Squared Error (FMSE)

4. Experiments and Discussion

✓ Standard model v.s. robust model

- From the visual sense



4. Experiments and Discussion

✓ Standard model v.s. robust model

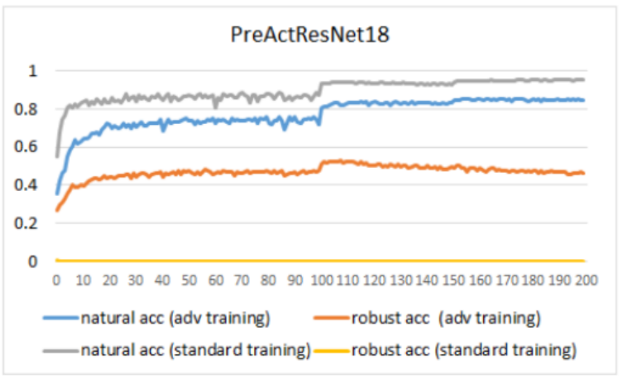
- From the evaluation metrics

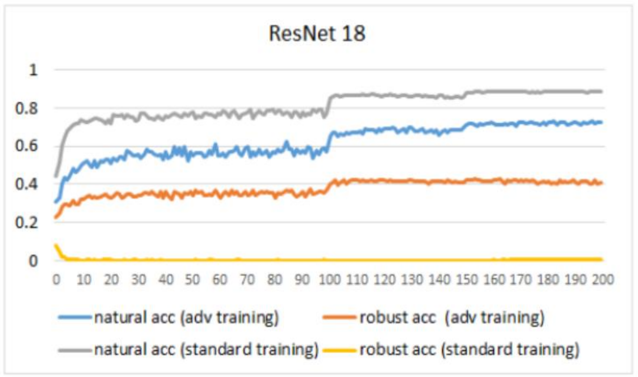
  - Adversarial training may pose a more serious risk of privacy leakage

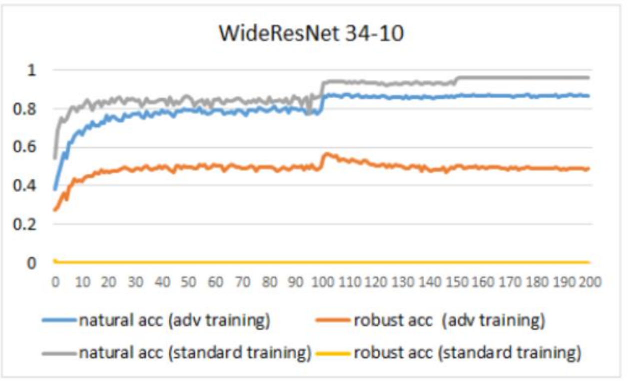
Arch	PreActResNet 18		ResNet 18		WideResNet34-10		SmallCNN	
Mode	Std.	Adv.	Std.	Adv.	Std.	Adv.	Std.	Adv.
PSNR	1.47	15.78	5.02	11.72	0.53	13.58	4.83	14.11
MSE	1.76	0.05	0.25	0.07	1.318	0.020	0.328	0.038
FMSE	1.31*e-01	4.02*e-04	5.58*e-05	1.72*e-07	4.52*e-01	9.20*e-05	1.75*e+02	7.26*e-04

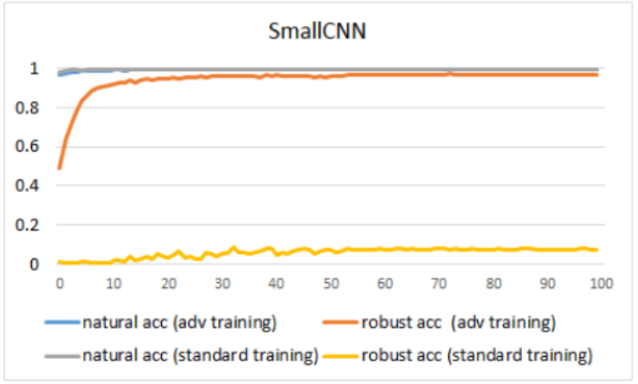
4. Experiments and Discussion

✓ Evaluation on training models





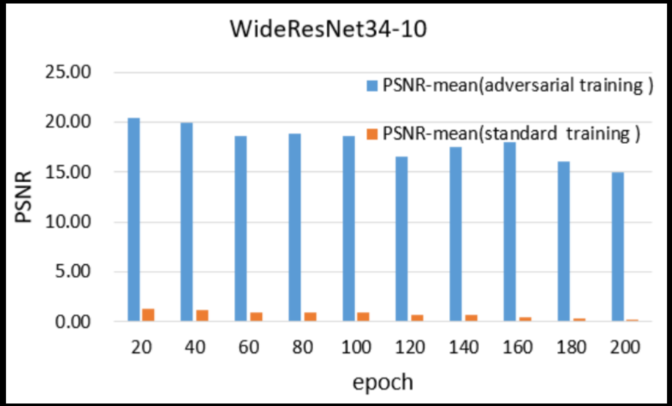


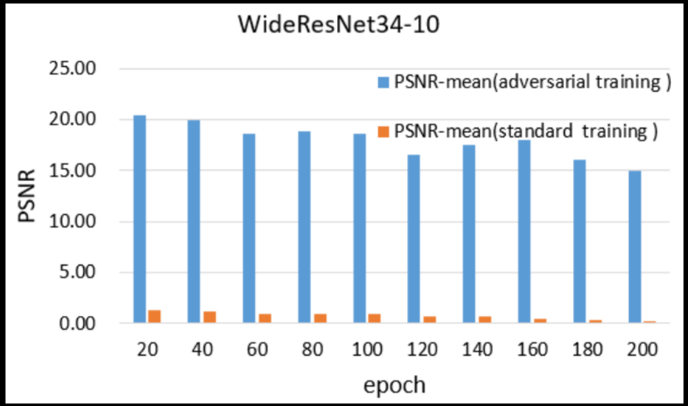


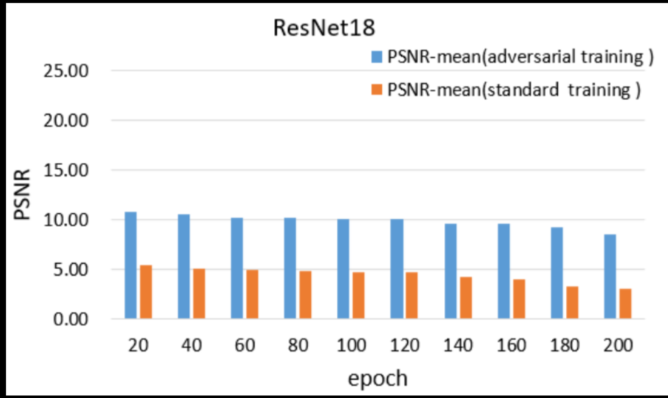
The evaluation on training models, two training modes are adopted in each type of model. x-axis represents the number of training epochs, y-axis represents the prediction accuracy.

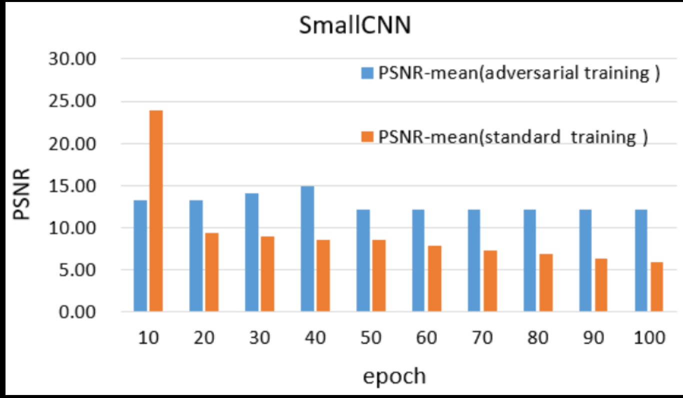
4. Experiments and Discussion

✓ Evaluation on different stage





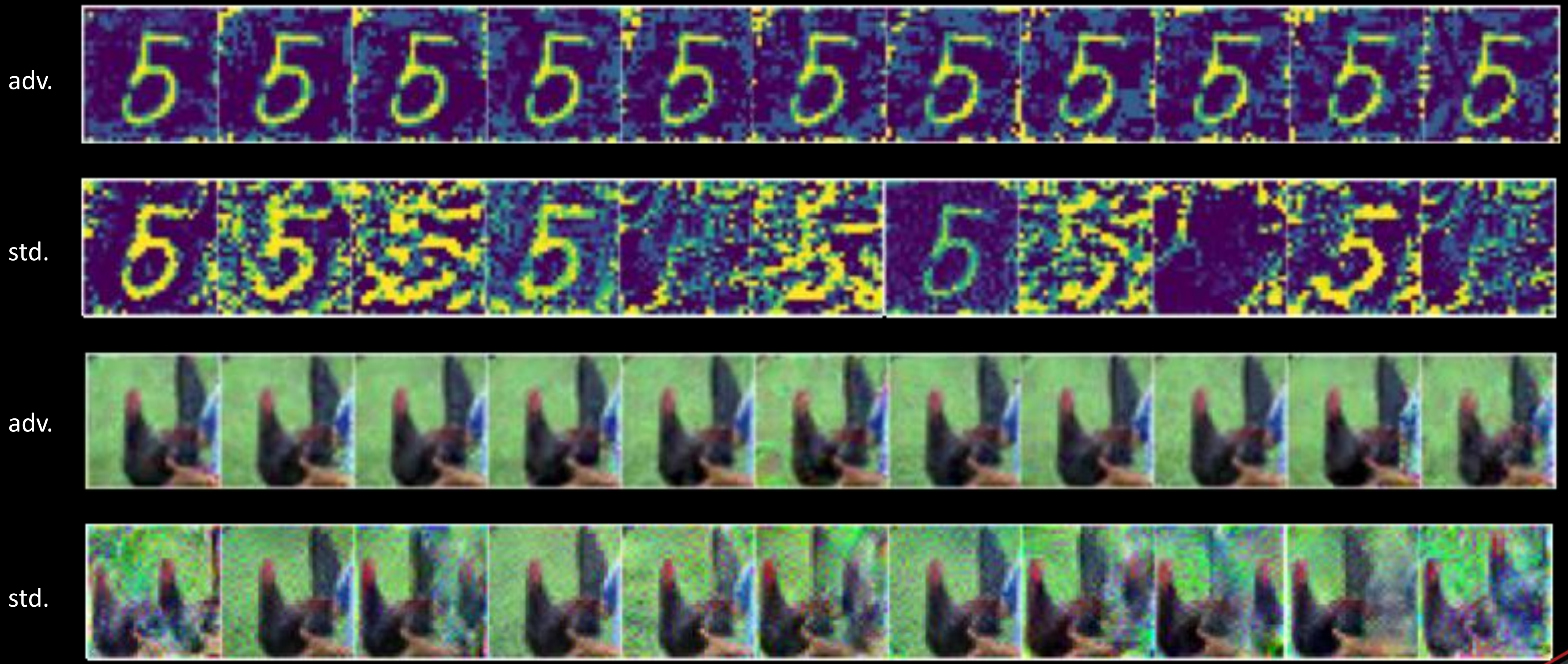




Robust models are more vulnerable to privacy attacks is universal and not a conclusion reached by chance.

4. Experiments and Discussion

✓ **Evaluation on reconstruction stability**



4. Experiments and Discussion

✓ **Trade-off between robustness and privacy**

model	train ε	train step_size	PSNR	MSE	FMSE
PreActResNet18	0/255	0/255	1.47	1.76	1.31*e-01
	4/255	1/255	14.56	0.06	6.81*e-04
	8/255	2/255	15.78	0.05	4.02*e-04
	16/255	4/255	16.43	0.02	4.97*e-06
	32/255	8/255	1.28	1.55	5.19*e-06
SmallCNN	0	0	4.83	0.32	1.75*e+02
	0.15	0.005	5.49	0.28	6.65*e+00
	0.30	0.01	11.04	0.03	7.26*e-04
	0.60	0.02	18.53	0.01	9.24*e-05
	0.90	0.03	7.35	0.19	4.07*e-02

- When becomes large enough (ε=32/255 for PreActResNet18 and ε=0.9 for SmallCNN), the data reconstruction quality will decreases rapidly.
- The robustness of the model and the risk of privacy leakage are not a simple positive or negative correlation

4. Discussion

✓ **Possible defenses**

- Differential privacy
- Homomorphic encryption
- Combine standard training and adversarial training

5. Conclusion

- Security is still one of the biggest challenges in deploying AI systems
- There are three most significant security challenges of AI

  - Vulnerabilities in AI components

  - New attacks targeting AI systems

  - The risk of AI abuse
- Training data of AI model can be stolen according to the gradient

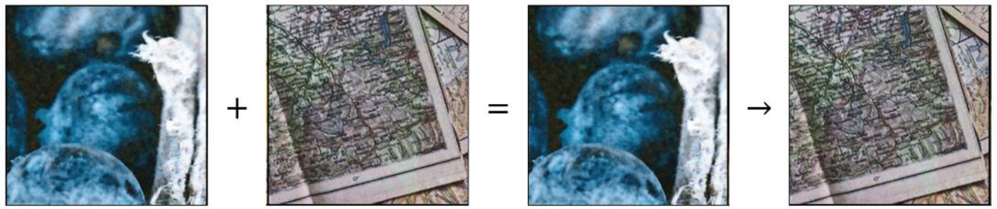
  - Model robustness will hurt data privacy
- Ultimate goals:

  - To build secure and trustworthy AI systems

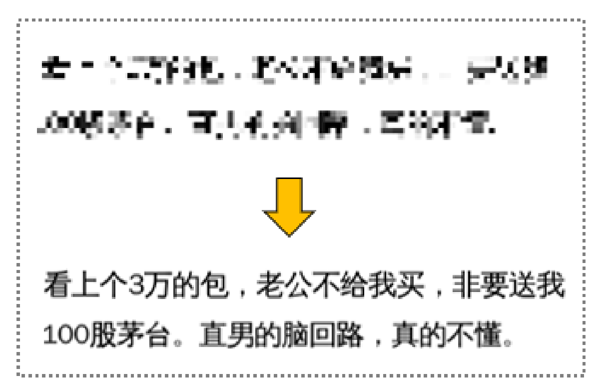
6. Appendix: Other Interesting Study

Steganography

Information hiding is one of the important ways to ensure data. We try to implement a case that image hiding in another image.



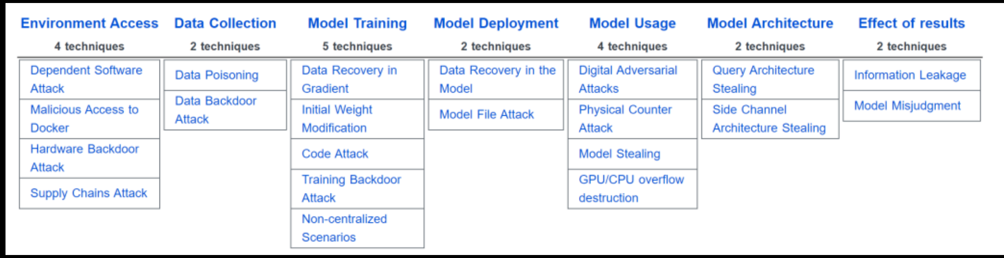
Mosaic Recovery



Use Seq2Seq model to restore mosaic text.

AISecMatrix

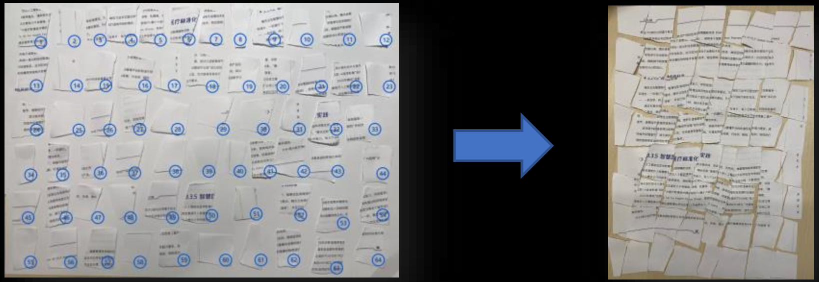
Based on ATT&CK paradigm, to provide developers and users a better guidance on the security problems of AI systems. https://github.com/AISecMatrix



Environment Access 4 techniques	Data Collection 2 techniques	Model Training 5 techniques	Model Deployment 2 techniques	Model Usage 4 techniques	Model Architecture 2 techniques	Effect of results 2 techniques
Dependent Software Attack	Data Poisoning	Data Recovery in Gradient	Data Recovery in the Model	Digital Adversarial Attacks	Query Architecture Stealing	Information Leakage
Malicious Access to Docker	Data Backdoor Attack	Initial Weight Modification	Model File Attack	Physical Counter Attack	Side Channel Architecture Stealing	Model Misjudgment
Hardware Backdoor Attack		Code Attack		Model Stealing		
Supply Chains Attack		Training Backdoor Attack		GPU/CPU overflow destruction		
		Non-centralized Scenarios				

Shredded Document

Reconstruction of shredded text documents with metric learning.



Thank You!

Feel free to ask questions:

mengyuntang@tencent.com

Tencent Security Platform Dpt.

Tencent Zhuque Lab

HITB SECCONF AMSTERDAM - 2021