DeepIP: Deep Neural Network Intellectual Property Protection with Passports

Lixin Fan, Kam Woh Ng, Chee Seng Chan, Qiang Yang

Presenter: Kam Woh Ng*, University of Surrey

*Work was done while the presenter was working in WeBank, China and University of Malaya, Malaysia

Github Page: https://kamwoh.github.io/DeepIPR

How do we `protect` DNN?

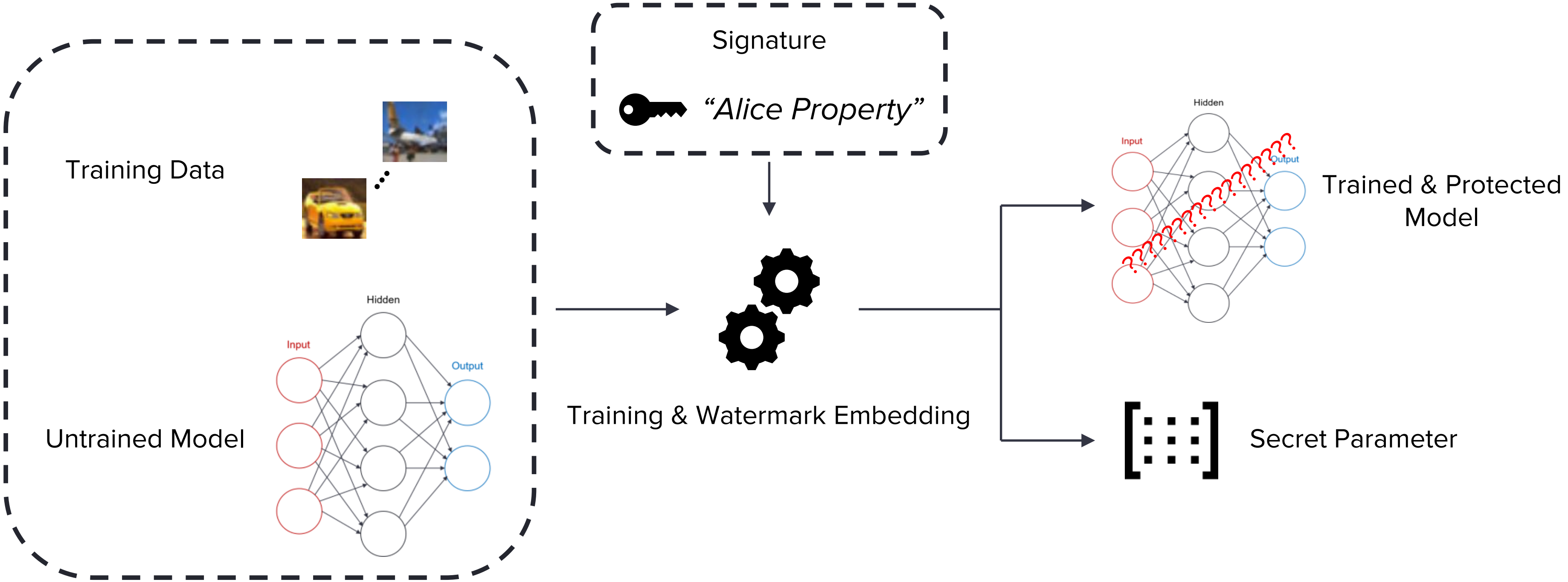
Conventional DNN watermarking methods

1. Feature based approach
   - Y. Uchida, Y. Nagai, S. Sakazawa, and S. Satoh, **“Embedding watermarks into deep neural networks” (2017)**
   - B. D. Rouhani, H. Chen, and F. Koushanfar, **“Deepsigns: A generic watermarking framework for IP protection of deep learning models” (2017)**

2. Trigger-set based approach
   - Yossi Adi, Carsten Baum, Moustapha Cisse, Benny Pinkas, and Joseph Keshet. **“Turning Your Weakness Into a Strength: Watermarking Deep Neural Networks by Backdooring” (2018)**
   - Jialong Zhang, Zhongshu Gu, Jiyong Jang, Hui Wu, Marc Ph. Stoecklin, Heqing Huang, and Ian Molloy. “**Protecting Intellectual Property of Deep Neural Networks with Watermarking” (2018)**

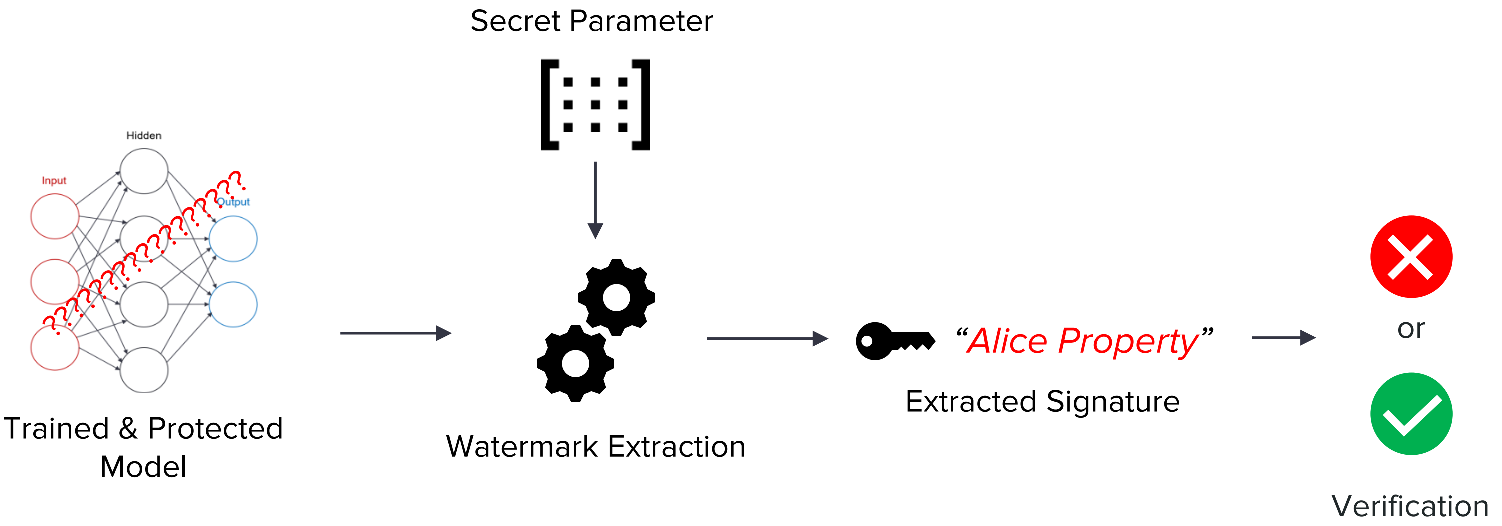
Feature-based approach (White-box)

Feature based watermark embedding



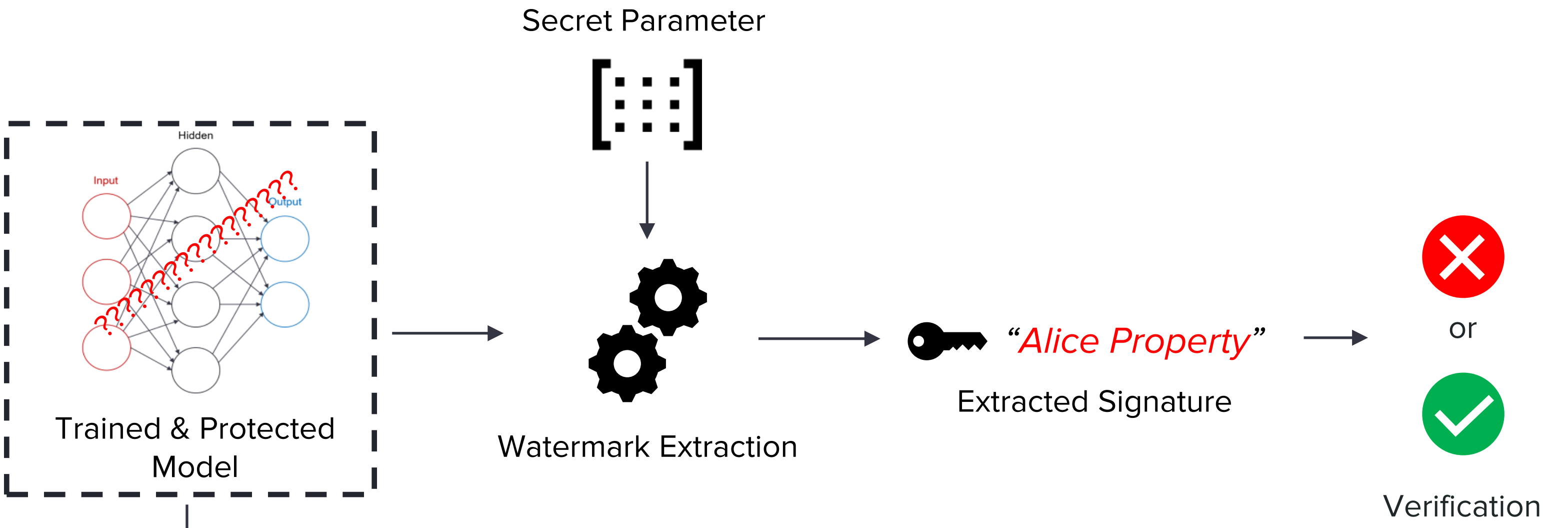
Feature-based approach (White-box)

Feature based watermark detection



Feature-based approach (White-box)

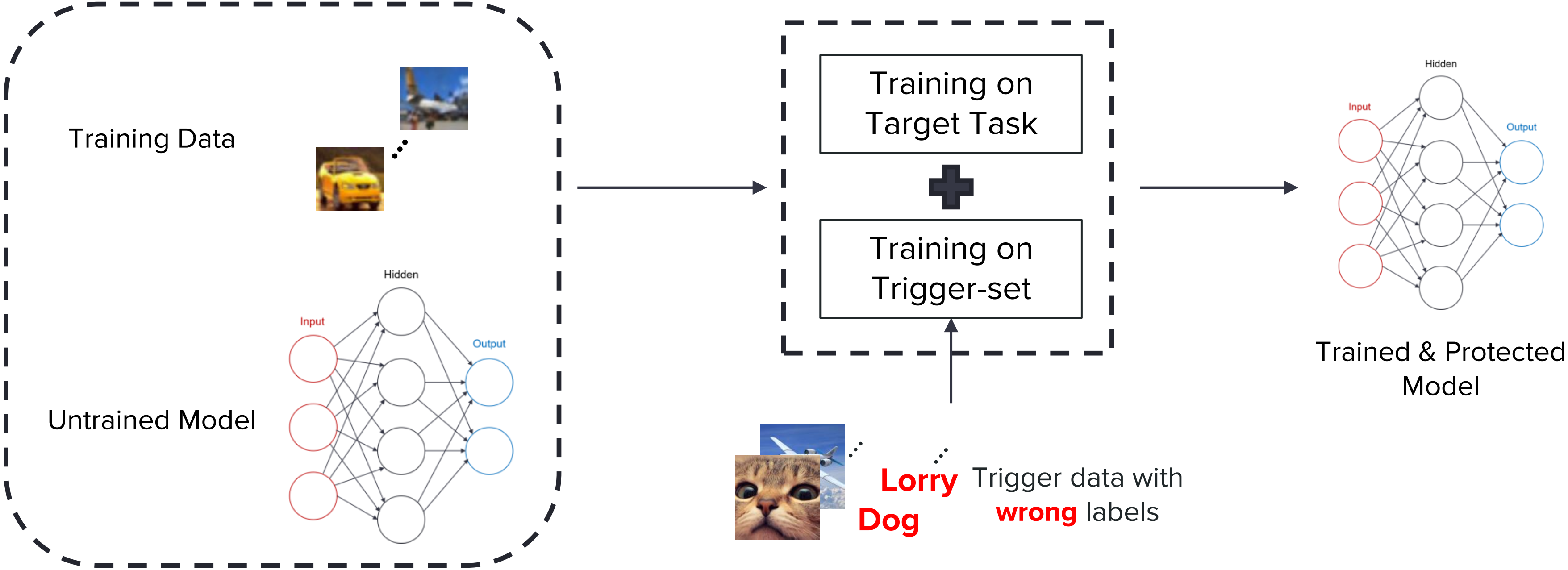
Feature based watermark detection



Disadvantage:
White-box ownership verification:
Need to access weights

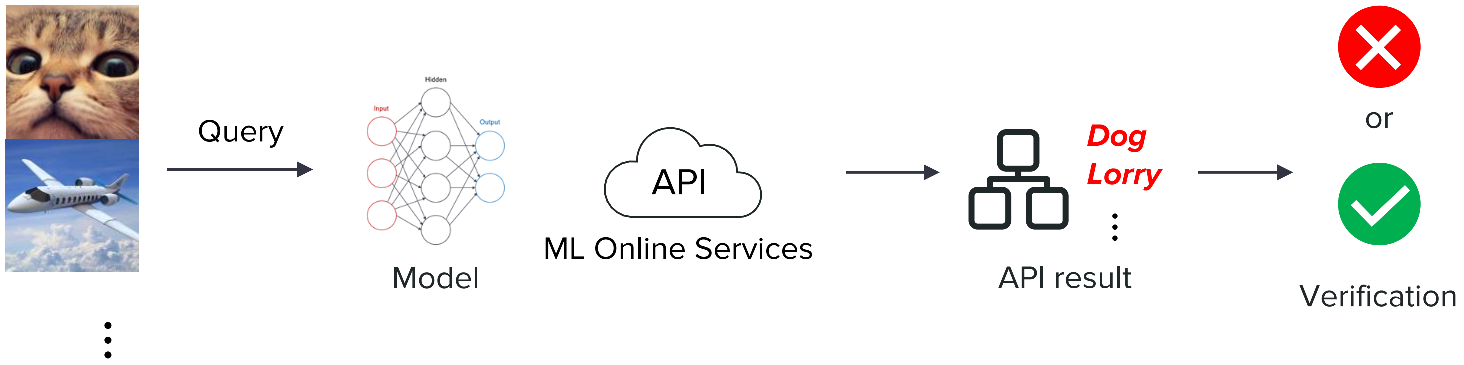
Trigger-set based approach (Black-box)

Trigger-set based watermark embedding



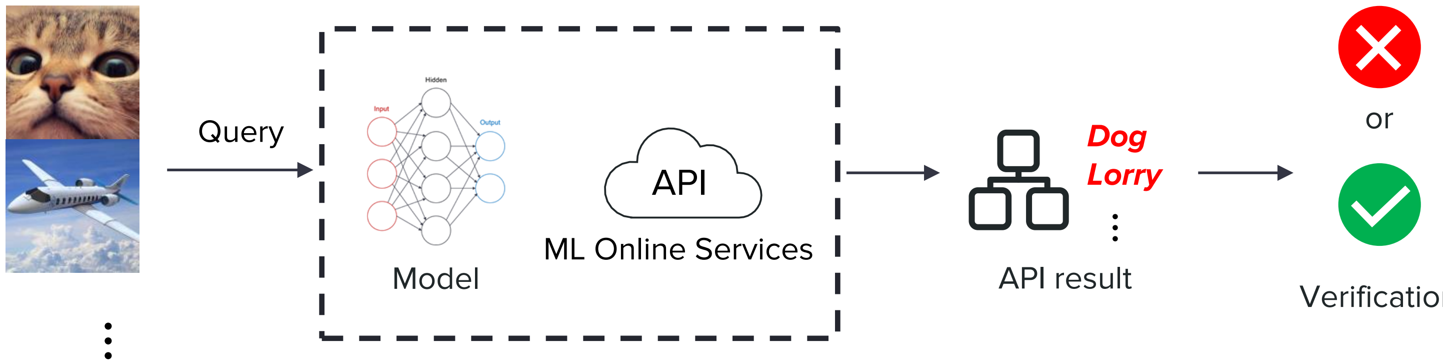
Trigger-set based approach (Black-box)

Trigger-set based watermark detection



Trigger-set based approach (Black-box)

Trigger-set based watermark detection

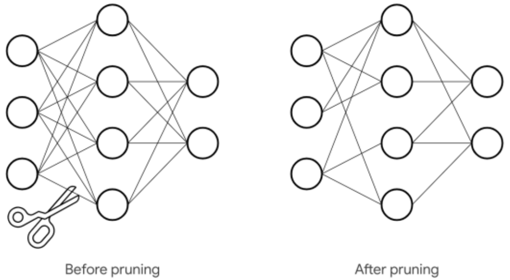


Black-box ownership verification

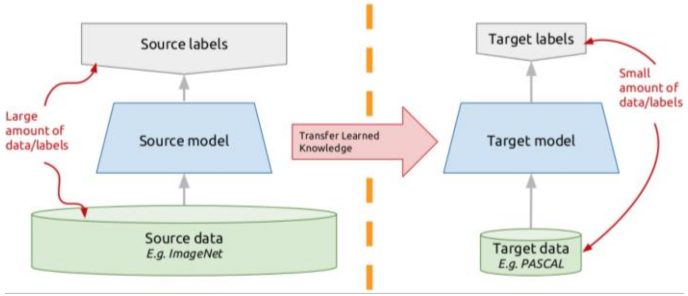
Can the watermarks be `attacked`?

Possible attacks to Ownership Protection

Watermarked Model

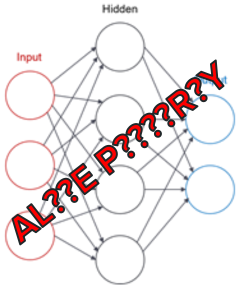


Model Pruning



Model Fine-tuning

Removal Attack



Model in Question

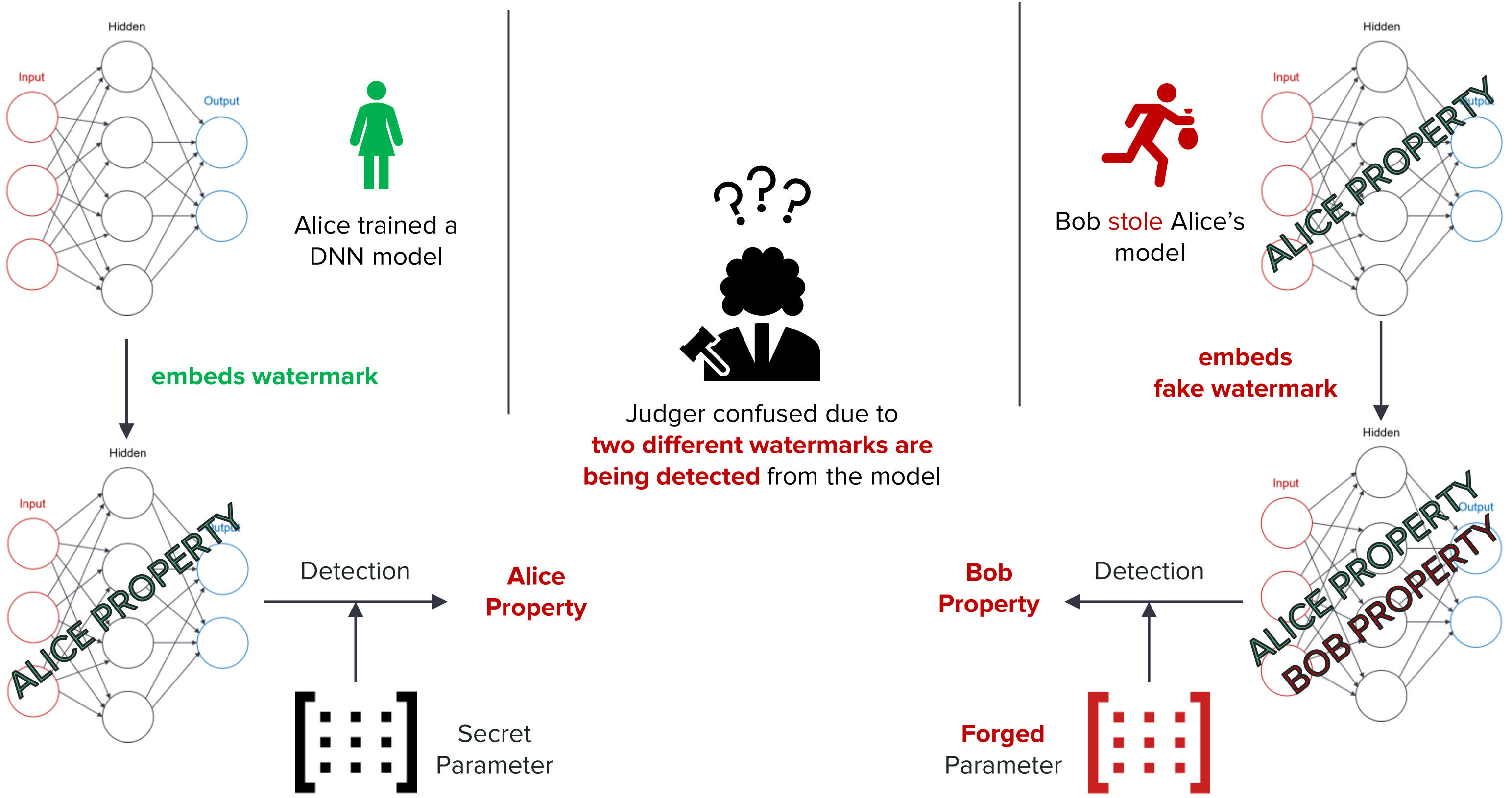
Effectiveness of Removal Attacks

- Watermark embedded in AlexNet for CIFAR10 classification

Removal Attacks	Feature based watermarking [1] (White-box)	Trigger-set based watermarking [2] (Black-box)
Model Pruning	**Strong** (100% watermark detected with 65% pruning rate)	**Strong** (100% watermark detected with 70% pruning rate)
Fine-tuning (CIFAR10 → CIFAR100)	**Strong** (100% watermark detected after fine-tuning)	**Weak** (25% watermark detected after fine-tuning)

Can we **forge** a watermark instead of removing it?

What is Ambiguity Attack?



Effectiveness of Ambiguity Attack

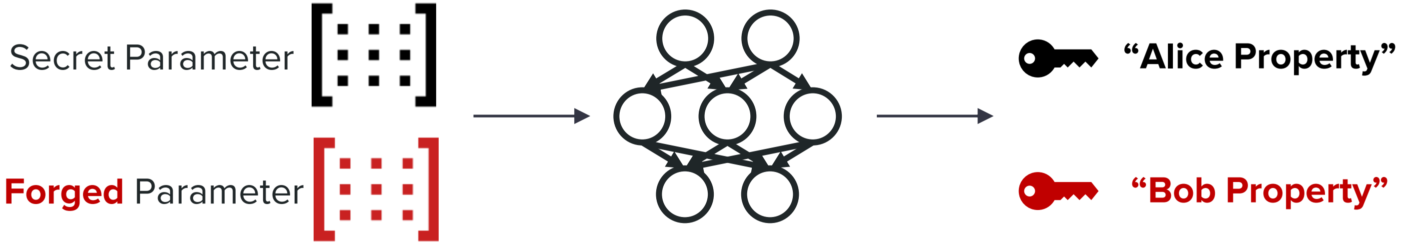
- Watermark embedded in AlexNet for CIFAR10 classification

Watermark approach	**Real Watermark**	**Fake Watermark**
Feature based (White-box)	100% **watermark** detected	100% **watermark** detected
Trigger-set based (Black-box)	100% **watermark** detected	100% **watermark** detected

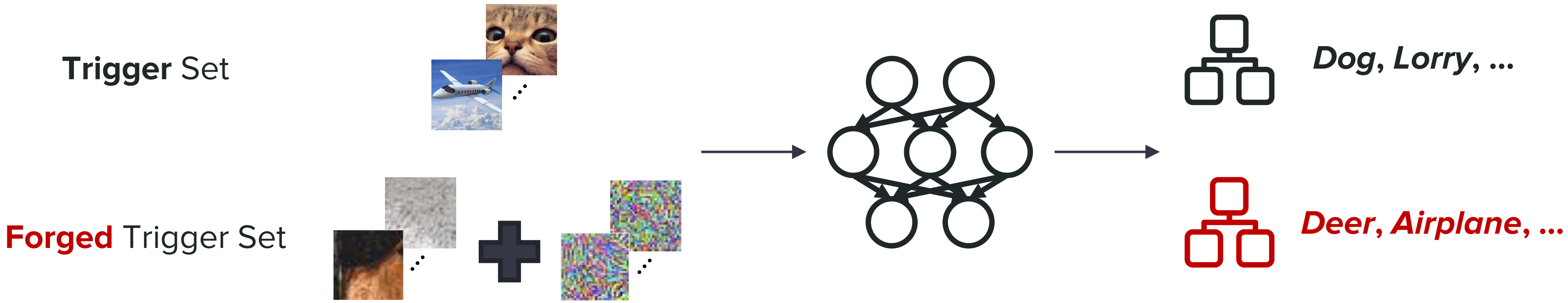
Watermark detection rate for both **real** and **fake** watermarks

Example of Ambiguity Attack

Feature based approach: Only train the **forged** parameter with the **frozen** model

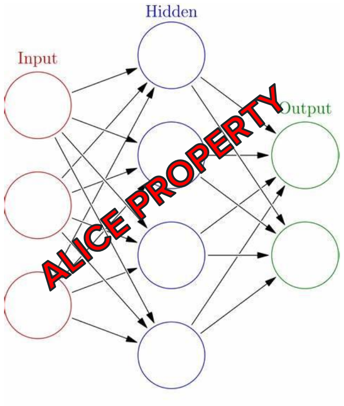


Trigger-set based approach: Train **an adversarial noise** on the **forged** trigger set

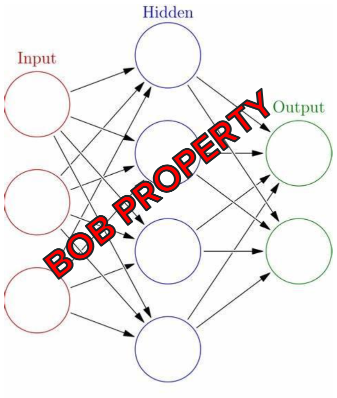


How to deal with `ambiguity` attack?

Current Situation

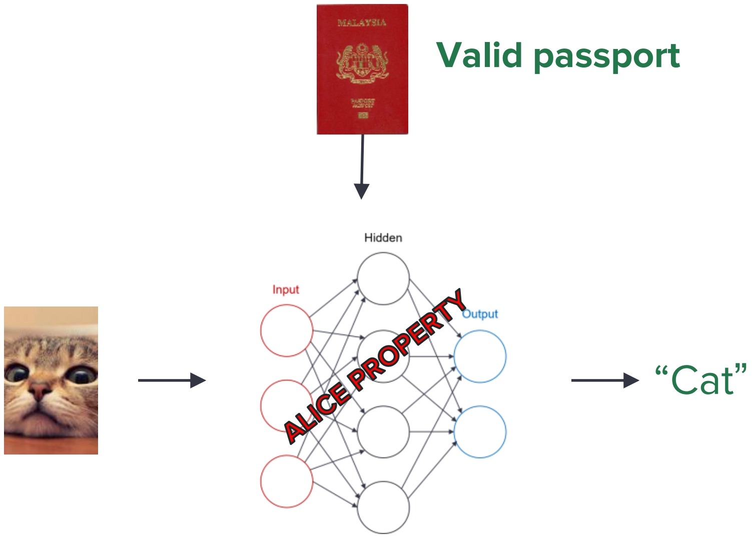


Protected Model
with original watermark

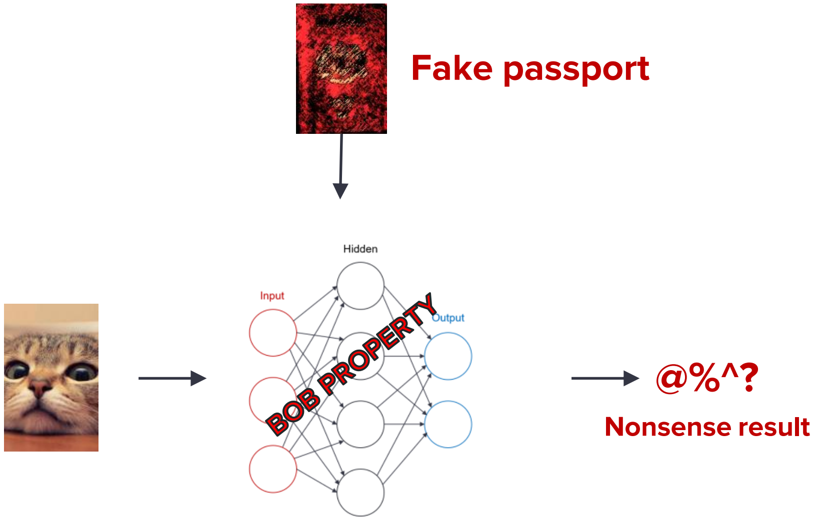


Copied Model
with fake watermark

Proposed Solution



Passport-protected Model
with **valid passport**

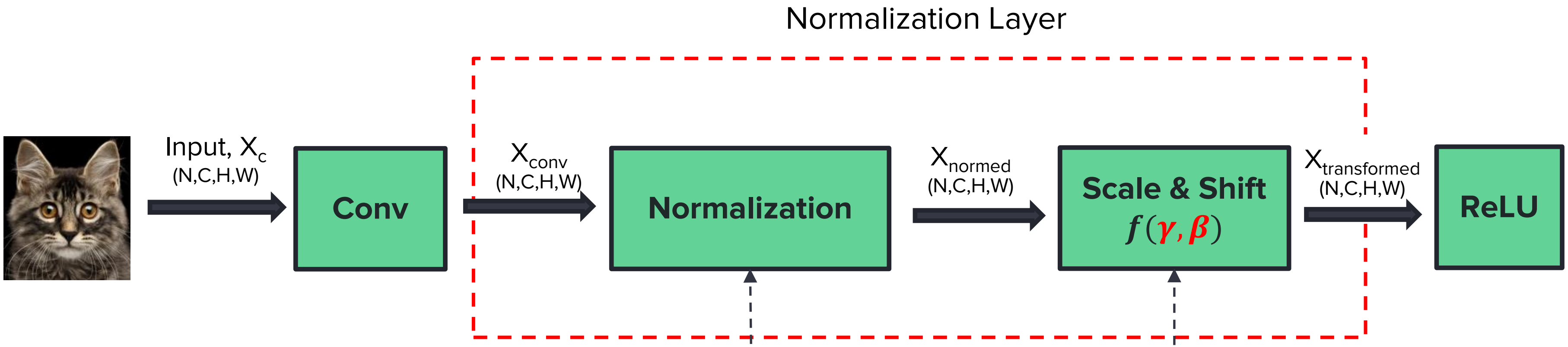


Passport-protected Model
with **fake passport**

Aim:

- Model cannot function without the **unique** and **valid** passport

Conventional Convolution Layer

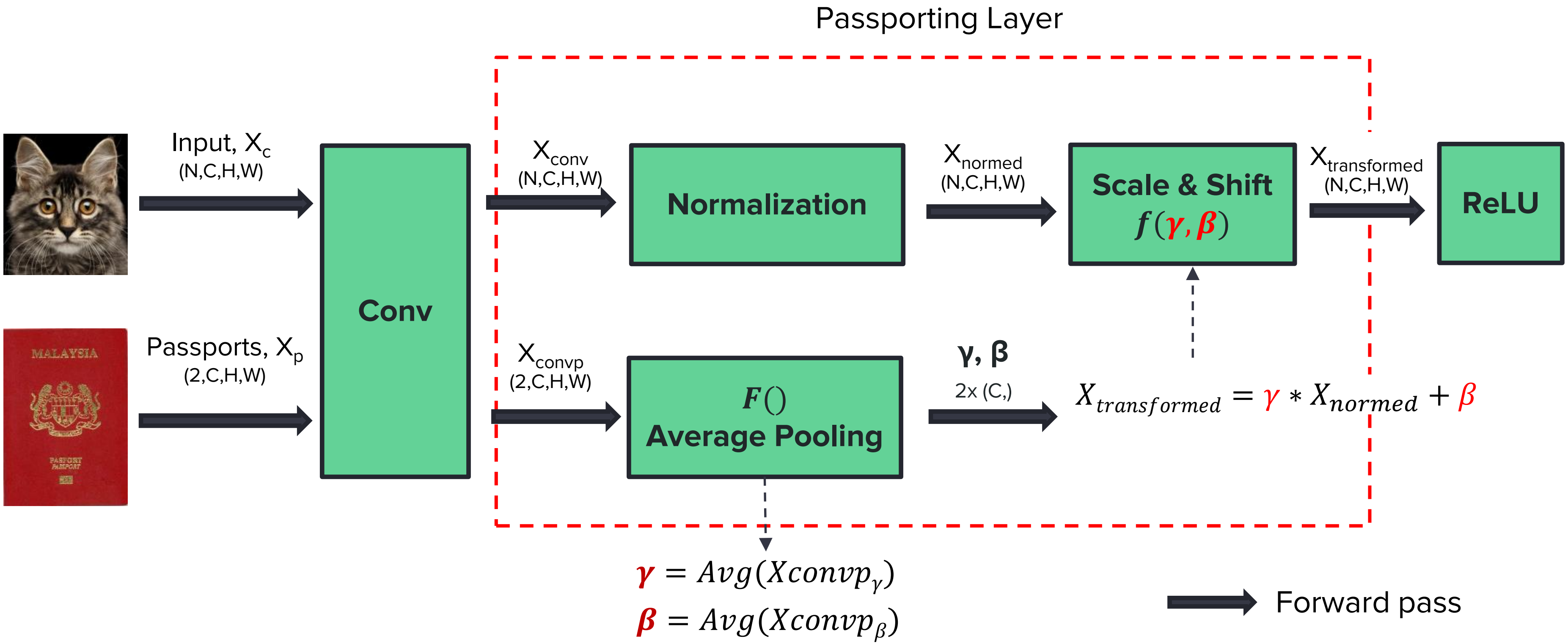


$$X_{normed} = \frac{X_{conv} - \mu}{\sigma^2}$$

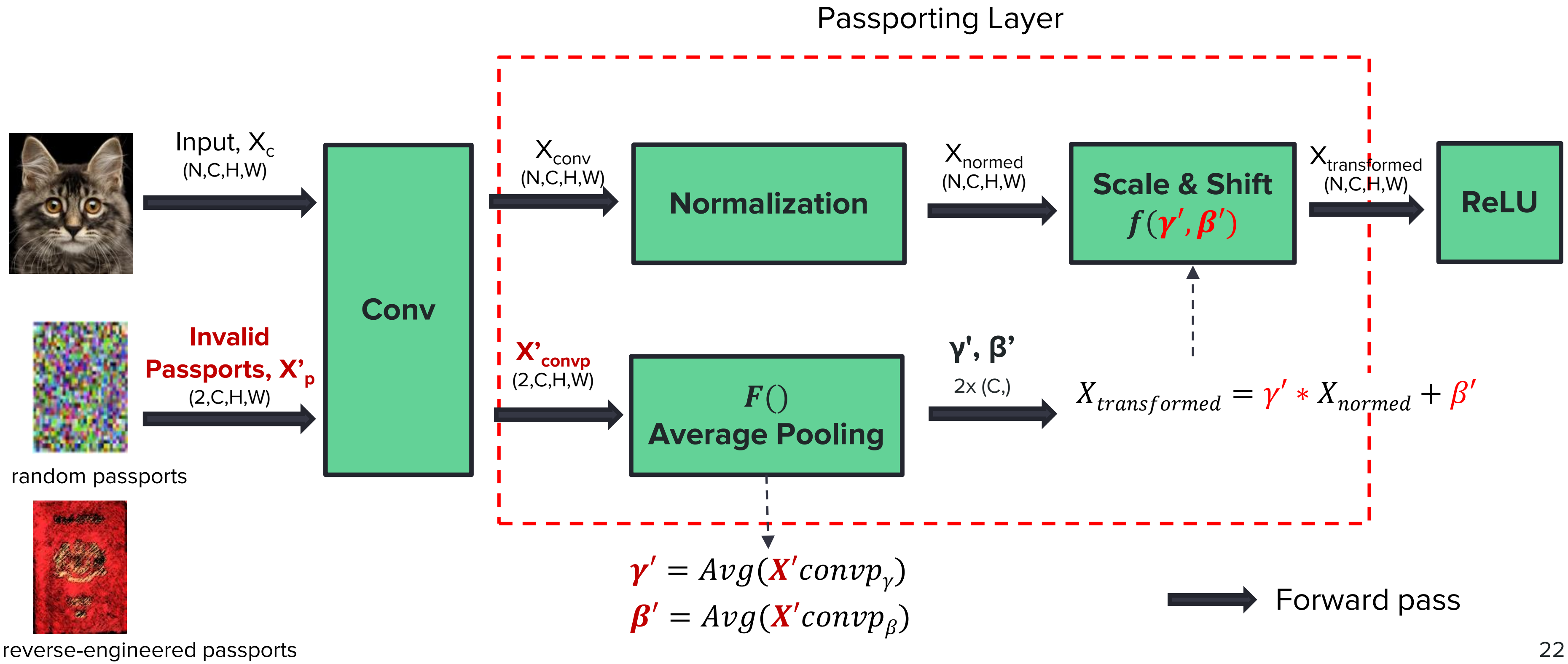
$$X_{transformed} = \gamma * X_{normed} + \beta$$

Forward pass

Passporting Layer



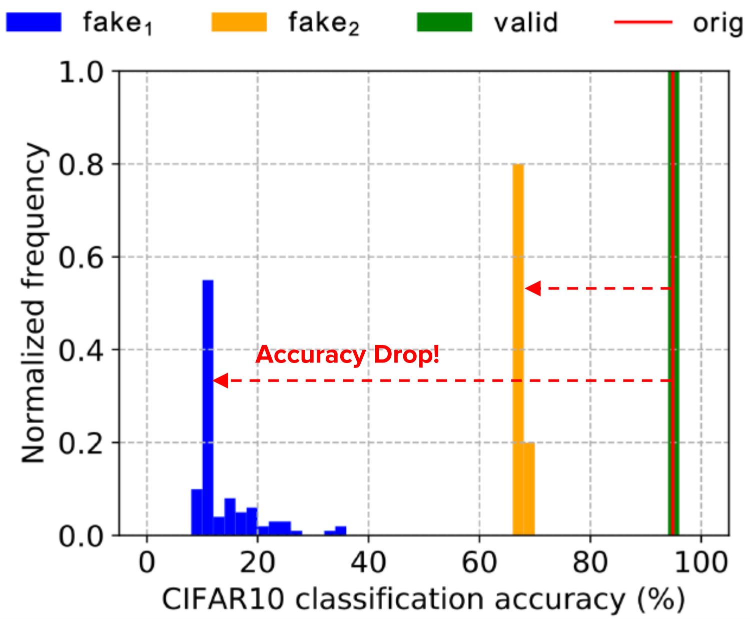
Passporting Layer



Effectiveness of Passport Protection

<u>Result of Invalid passports</u>

Ambiguity attack	Effect
$Fake_1$ (random passports) 	**Random guessing** (at max 35%)
$Fake_2$ (reverse-engineered passports) 	**Performance deteriorated** (at max 70%)



Example of $ResNet_p$-18 performance on CIFAR10 when performing different ambiguity attacks ($fake_1$ & $fake_2$)

Embedding Binary Signatures by Sign of Scale Factors (Gamma)

Enforce scale factor to take **either positive or negative signs** as designated

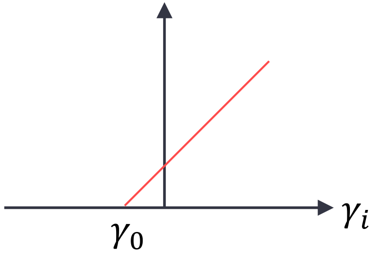
Using hinge-loss like of regularization: **Sign-Loss**

64 channels can embed 8 bytes signature

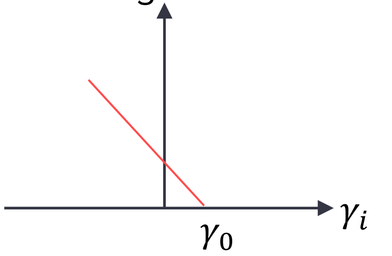
0.1 as default

Sign Loss = $\sum_{i=1}^{C} \max(\gamma_0 - \gamma_i \boxed{b_i}, 0)$

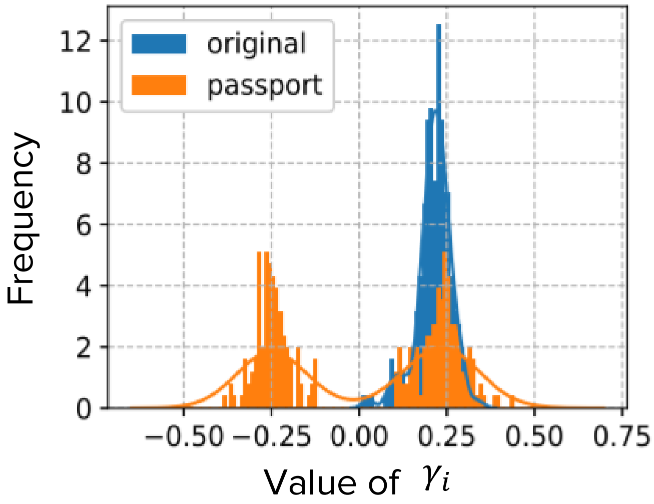
b: [-1 1 -1 1 ...]

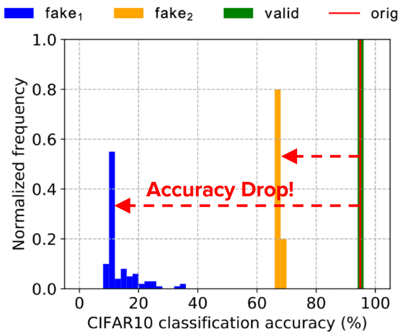


If b is -1 then γ_i need to be negative



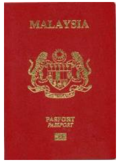
If b is 1 then γ_i need to be positive





Summary of Ambiguity Attacks

Summarized result done on: AlexNet & ResNet18
Datasets: CIFAR10, CIFAR100, ImageNet

Ambiguity Attacks		Inference Phase	Verification Phase
	Fake_1, Random Passport	**- Random Guessing** **- Useless Model**	**- Useless Infringement**
	Fake_2, Reverse-Engineered Passport	**- Deteriorated Performance** **- Useless Model**	**- Useless Infringement**
	Fake_3, Copied Passport	**- Performance Detained** **- Signature Detected**	**- Ownership Verified**

Take Home message

- **Protection on DNN** is urgently needed!

- Some **existing** watermarking approaches are **vulnerable to ambiguity attack**

- **Passport-based approach** provided better protection in terms of **robustness against removal attack (non-removable) and ambiguity attack (unique signature)**

- **Passport-protected DNN model** will **only perform well if and only if a valid passport is used**, else the performance will be significantly deteriorated

More Details and Implementation

Project Page: https://kamwoh.github.io/DeepIPR

GitHub: https://github.com/kamwoh/DeepIPR

Contact: kamwoh@gmail.com

References

[1] Yusuke Uchida, Yuki Nagai, Shigeyuki Sakazawa, and Shin'ichi Satoh. Embedding watermarks into deep neural networks. In Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, pages 269–277, 2017

[2] Y Adi, C Baum, M Cisse, B Pinkas, and J Keshet. Turning your weakness into a strength: Watermarking deep neural networks by backdooring. In 27th USENIX Security Symposium (USENIX), 2018.

[3] Jialong Zhang, Zhongshu Gu, Jiyong Jang, Hui Wu, Marc Ph Stoecklin, Heqing Huang, and Ian Molloy. Protecting intellectual property of deep neural networks with watermarking. In Proceedings of the 2018 on Asia Conference on Computer and Communications Security (ASIACCS), pages 159–172, 2018.

Thank You for Listening!