NeurIPS | 2019

Thirty-third Conference on Neural Information Processing Systems

WeBank 微众银行

UNIVERSITY OF MALAYA

Rethinking Deep Neural Network Ownership Verification: Embedding Passports to Defeat Ambiguity Attacks

Lixin Fan[1], Kam Woh Ng[1,2], Chee Seng Chan[2]

[1]WeBank AI Lab, Shenzhen, China

[2]Center of Image and Signal Processing, University of Malaya, Kuala Lumpur, Malaysia

{ lixinfan@webank.com ; jinhewu@webank.com ; cs.chan@um.edu.my }

研讨会官网：
扫描二维码

活动介绍链接：
扫描二维码

Rethinking Deep Neural Network Ownership Verification: Embedding Passports to Defeat Ambiguity Attacks

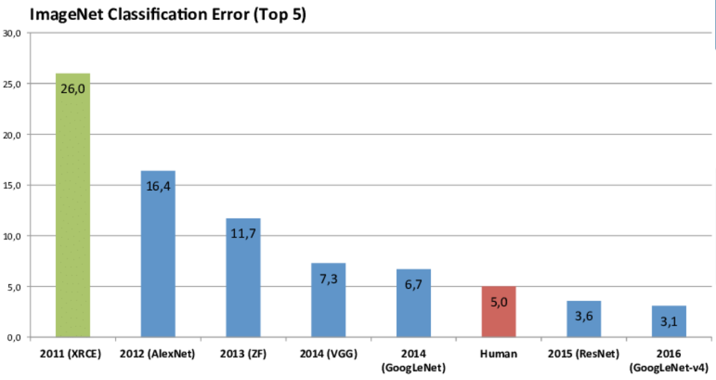
Paper: http://papers.nips.cc/paper/8719-rethinking-deep-neural-network-ownership-verification-embedding-passports-to-defeat-ambiguity-attacks

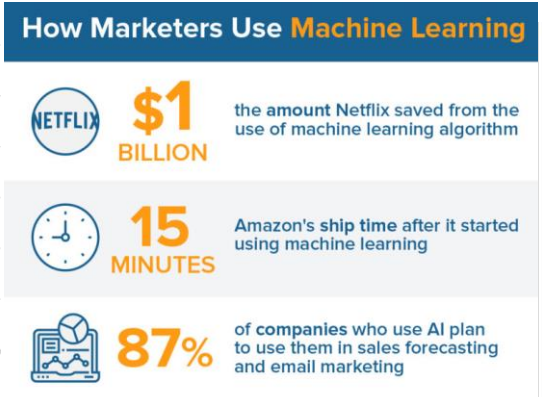
Arxiv: https://arxiv.org/abs/1909.07830

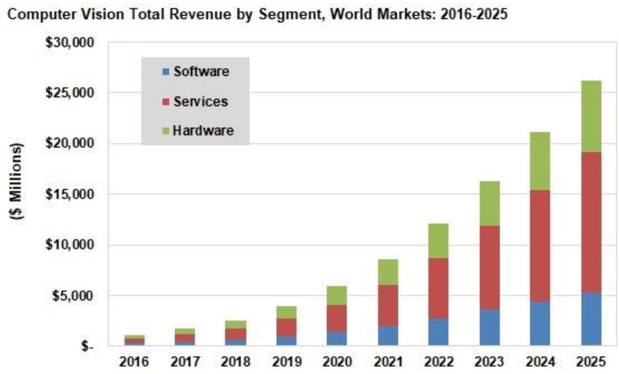
Github Page: https://kamwoh.github.io/DeepIPR

Machine Learning as a Service (MLaaS)

- More and more **business models and services** using Deep Neural Network (DNN)



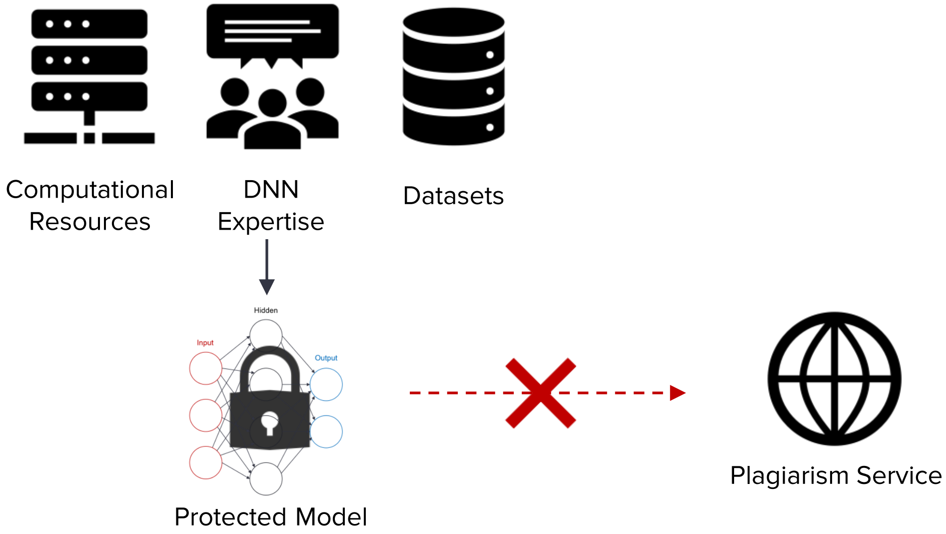




https://www.embedded-vision.com/industry-analysis/market-analysis/computer-vision-hardware-software-and-services-market-reach-262-bi
https://financesonline.com/machine-learning-statistics/

Protection on DNN model is needed!

- Companies **invested a lot** to create powerful models
- They are **easily copied** and used by plagiarizers.
- We **need a protection** on DNN **from being illegally copied, distributed and abused.**



How do we protect DNN?

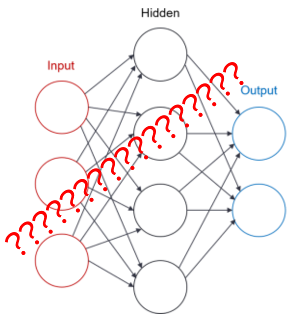
Conventional DNN watermarking methods

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

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"**
   - 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"**

Feature-based approach

Feature based watermark detection



Watermarked Model

Transformation Matrix

Watermark Extraction Process

"ALICE PROPERTY"

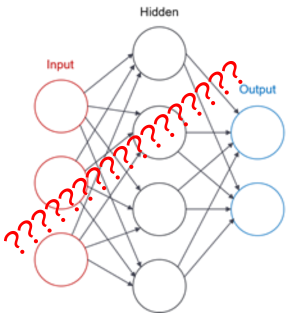
or

Verification

**Embedding process is identical

Feature-based approach

Feature based watermark detection



Watermarked Model

Transformation Matrix

Watermark Extraction Process

"ALICE PROPERTY"

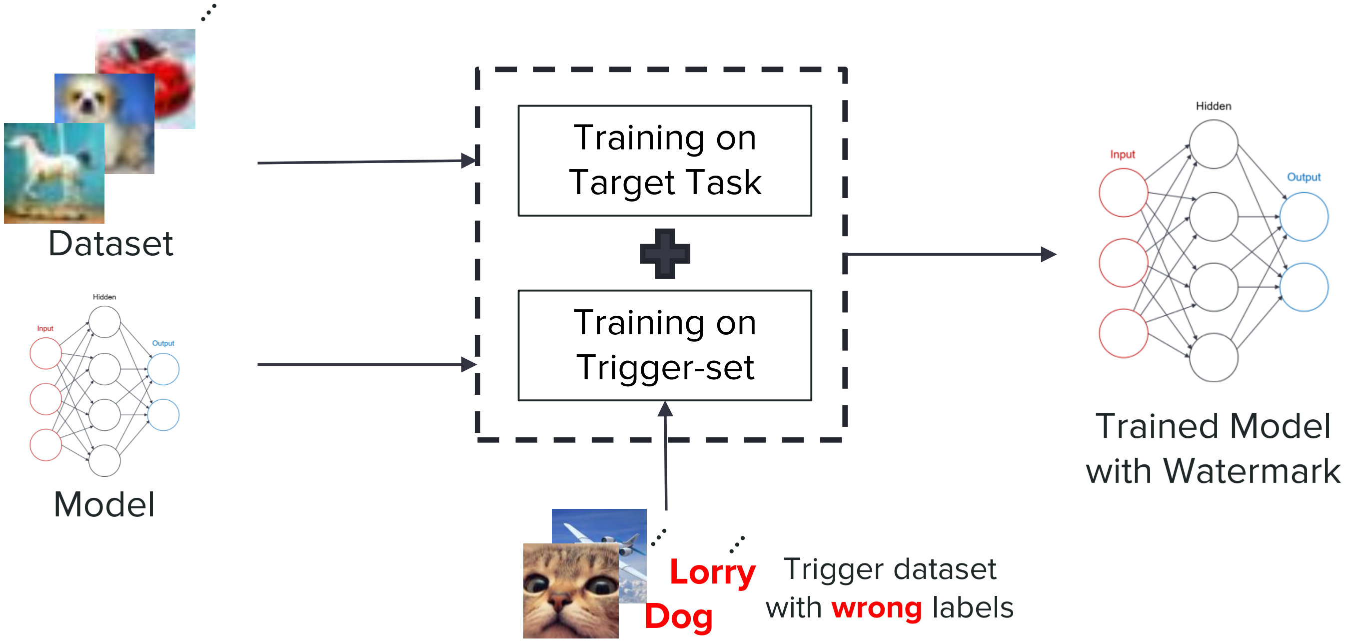
or

Verification

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

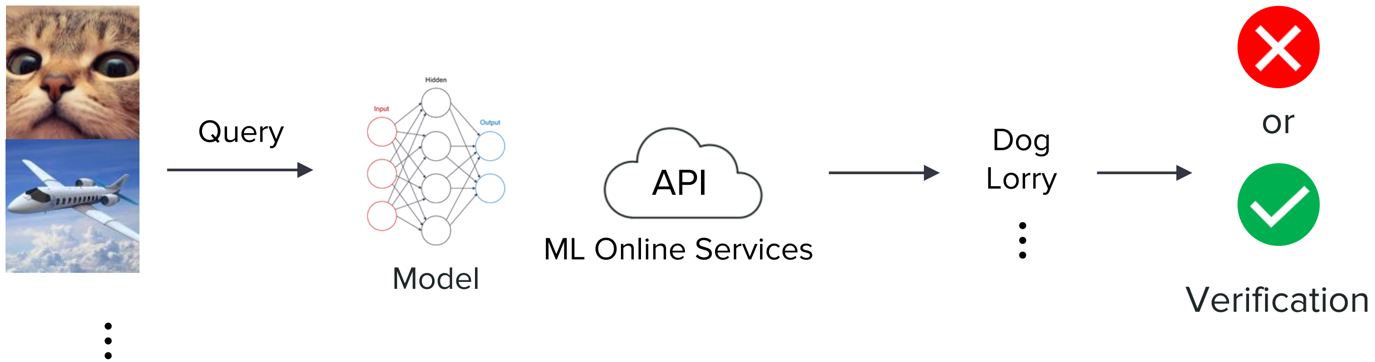
Trigger-set based approach

Trigger-set based watermark embedding



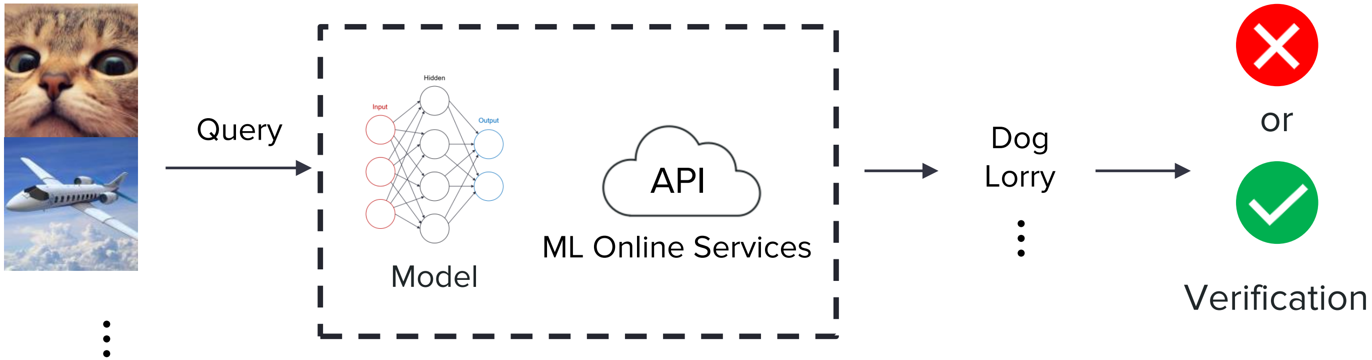
Trigger-set based approach

Trigger-set based watermark detection



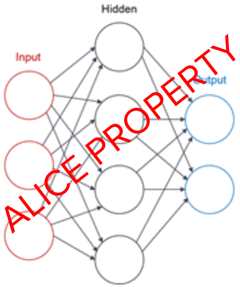
Trigger-set based approach

Trigger-set based watermark detection

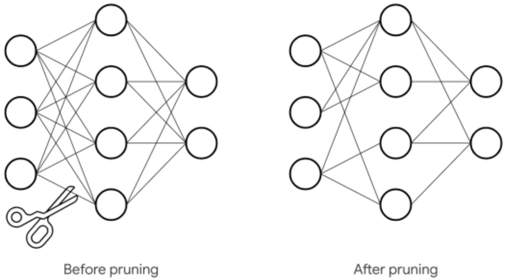


Black-box ownership verification

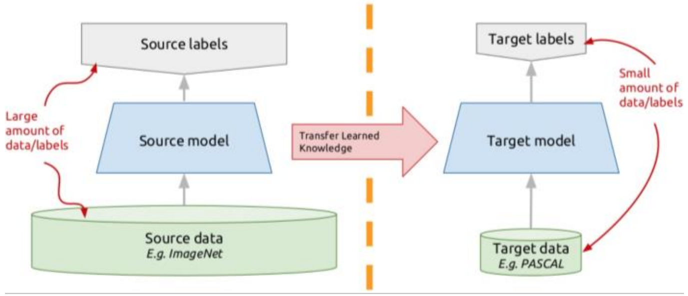
Possible attacks to Ownership Protection



Watermarked Model

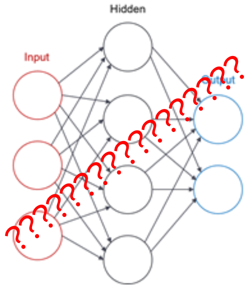


Model Pruning



Model Fine-tuning

Removal Attack



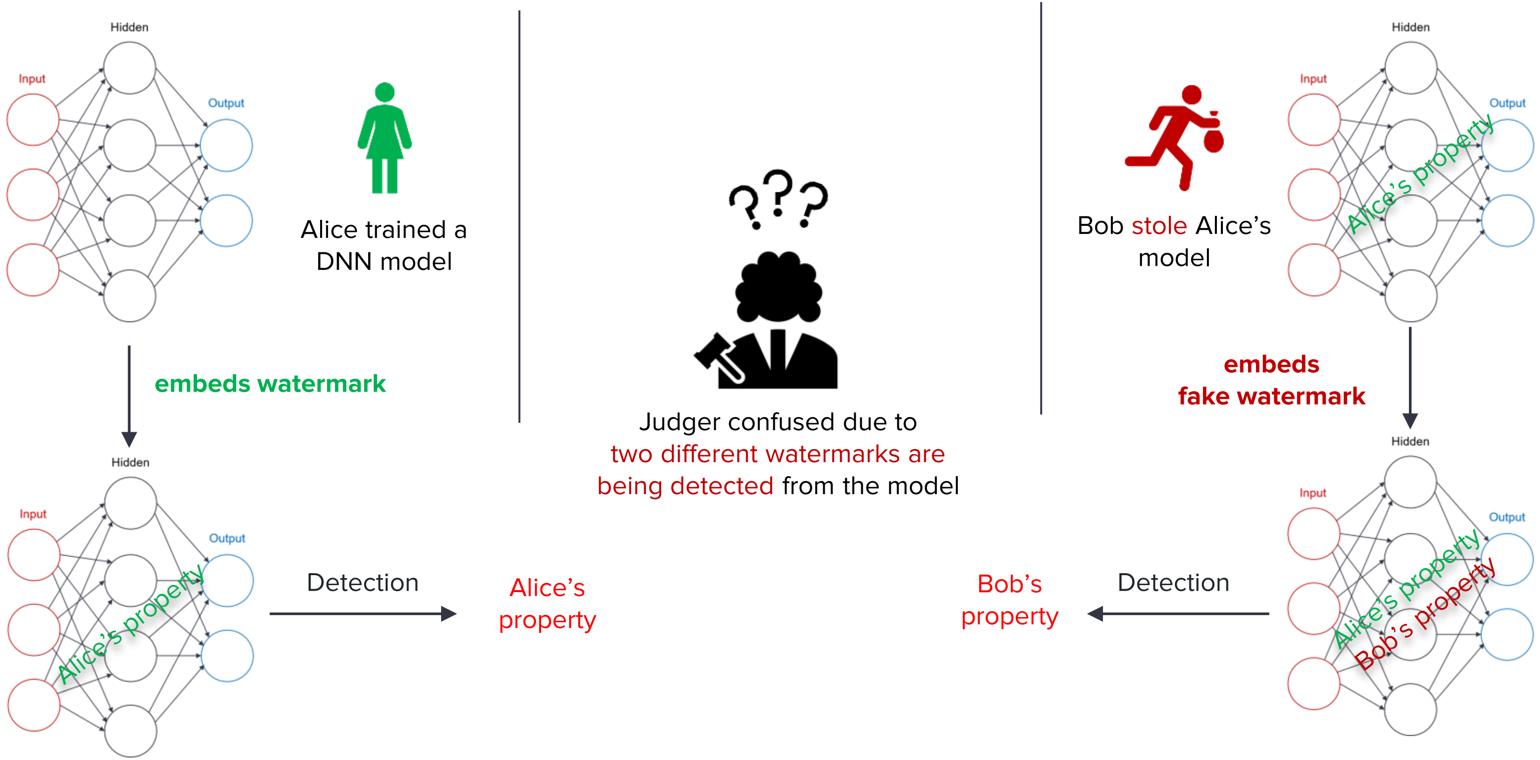
Model in Question

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)

What is Ambiguity Attack



Effectiveness of Ambiguity Attack

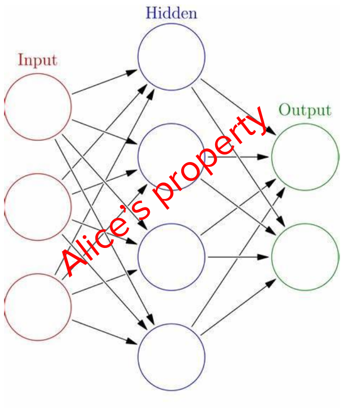
- Watermark embedded in AlexNet for CIFAR10 classification

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

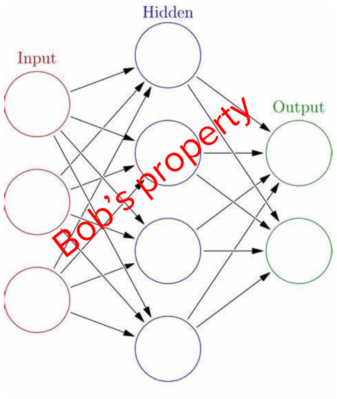
Watermark detection rate for both **real** and **fake** watermarks

How to deal with `ambiguity` attack?

Current Situation

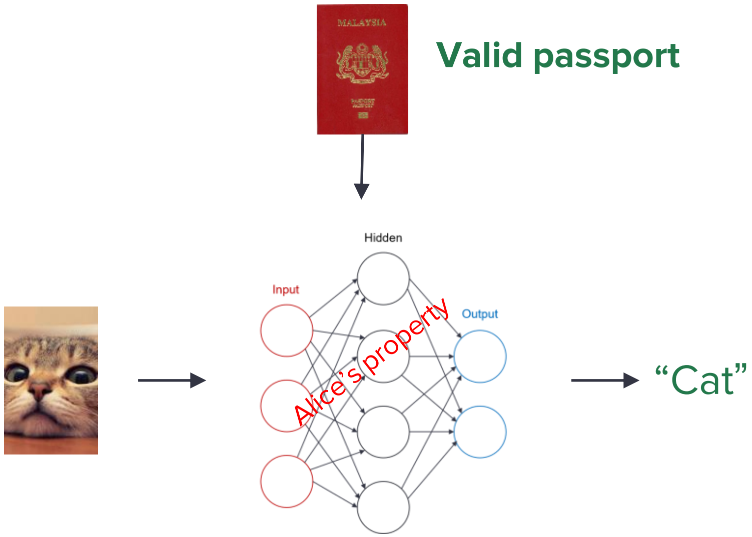


Passport-protected Model
with original watermark

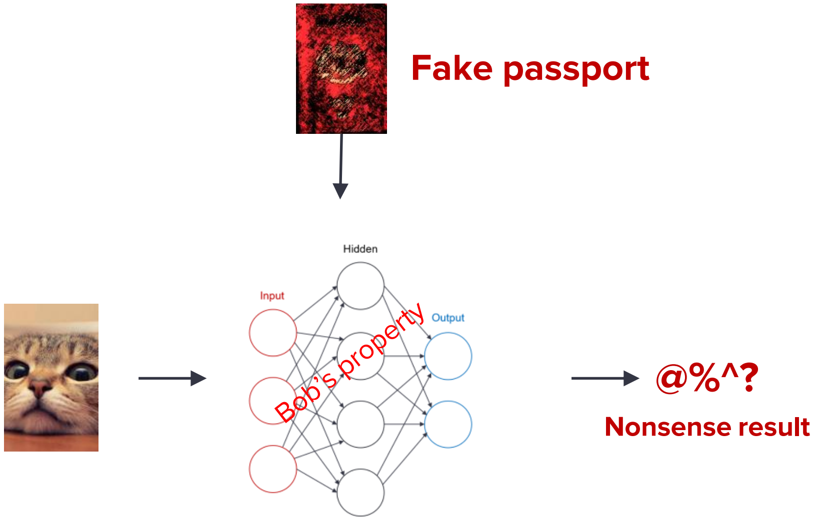


Copied Model
with fake watermark

Proposed Solution



Passport-protected Model
with **valid passport**

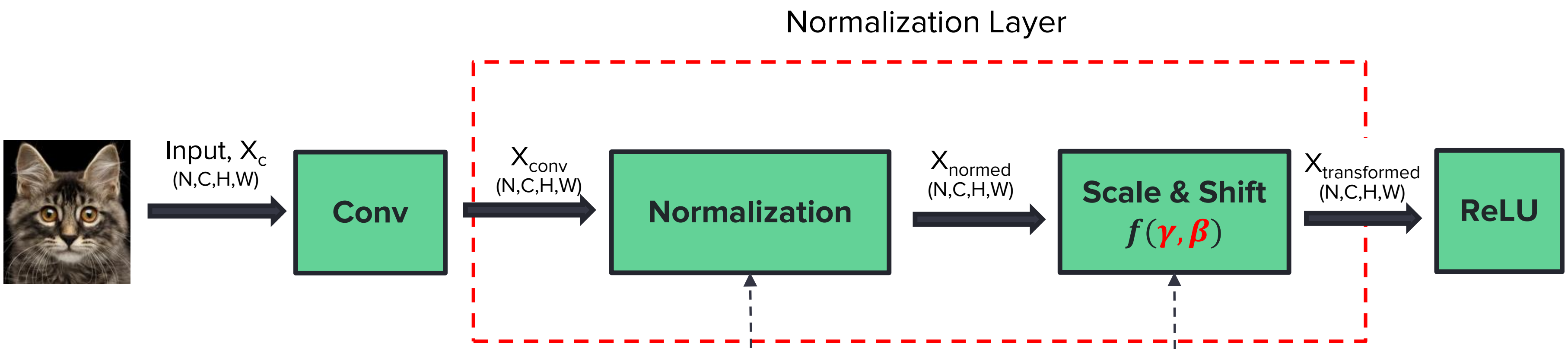


Passport-protected Model
with **fake passport**

Aim:

- Model cannot function without valid passport

Conventional Convolution Layer

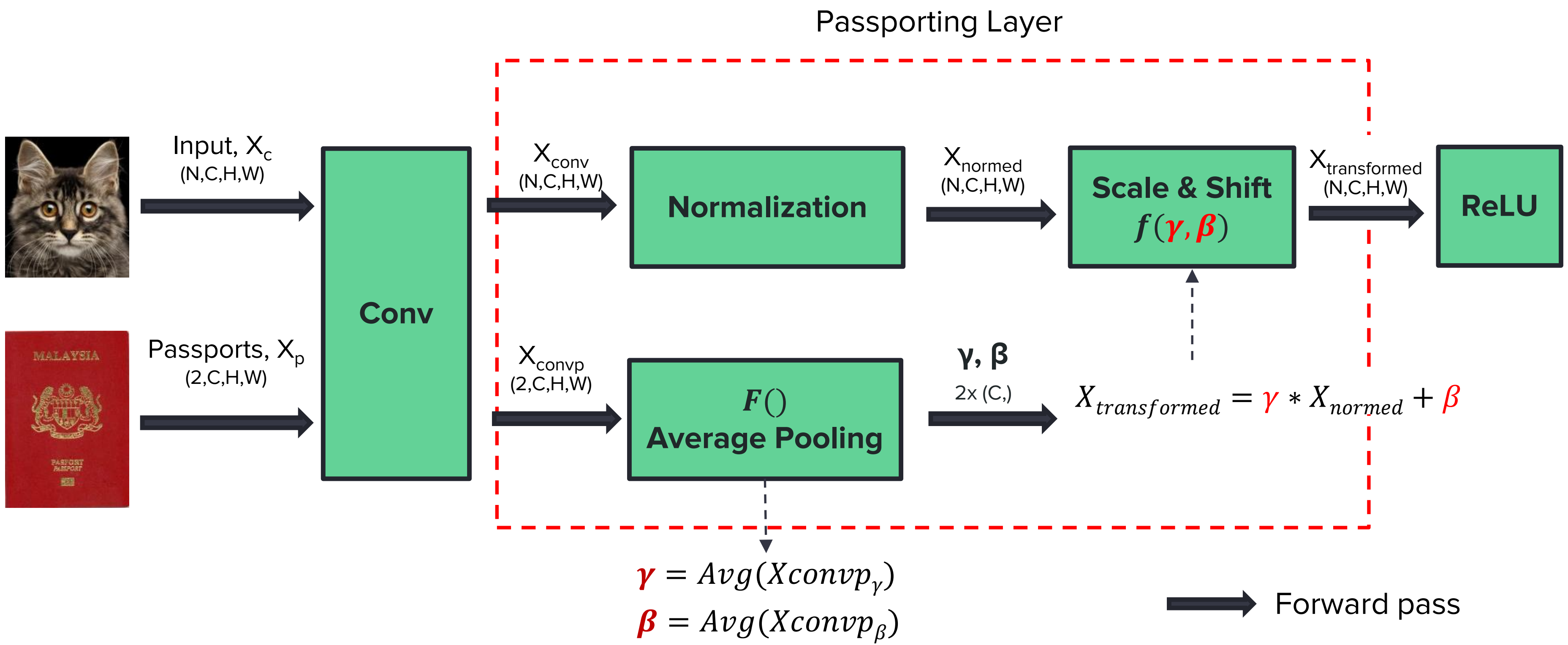


$$X_{normed} = \frac{X_{conv} - \mu}{\sigma^2}$$

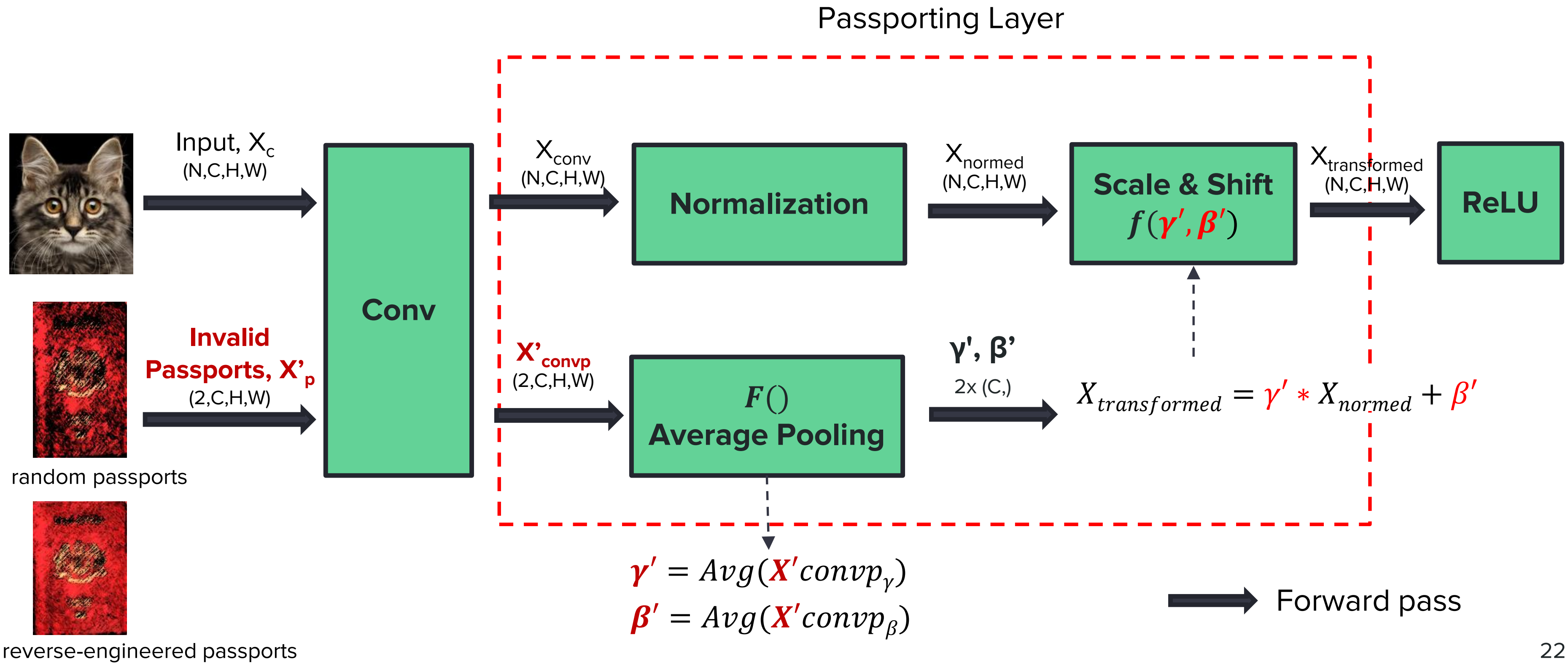
$$X_{transformed} = \gamma * X_{normed} + \beta$$

Forward pass

Passporting Layer



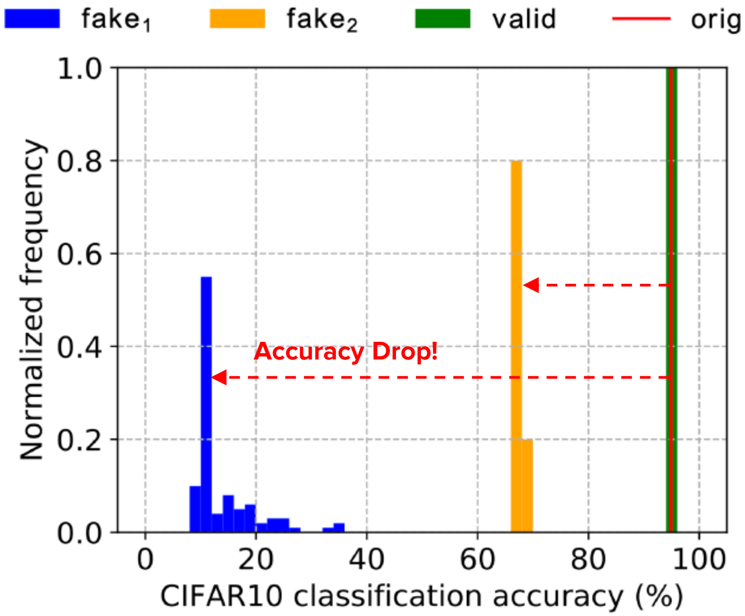
Passporting Layer



Effectiveness of Passport Protection

<u>Result of Invalid passports</u>

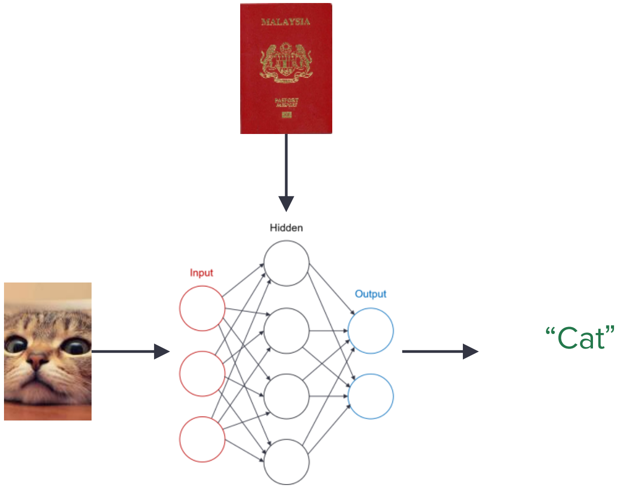
Ambiguity attack	Effect
Fake$_1$ (random passports) 	**Random guessing** (at max 35%)
Fake$_2$ (reverse-engineered passports) 	**Performance deteriorated** (at max 70%)



Example of ResNet$_p$-18 performance on CIFAR10 when performing different ambiguity attacks (fake$_1$ & fake$_2$)

Ownership Verification with Passports (Scheme 1)

Training & Inference



Passport is distributed with the trained DNN model

Model ownership is verified by **passports, performance and signature**

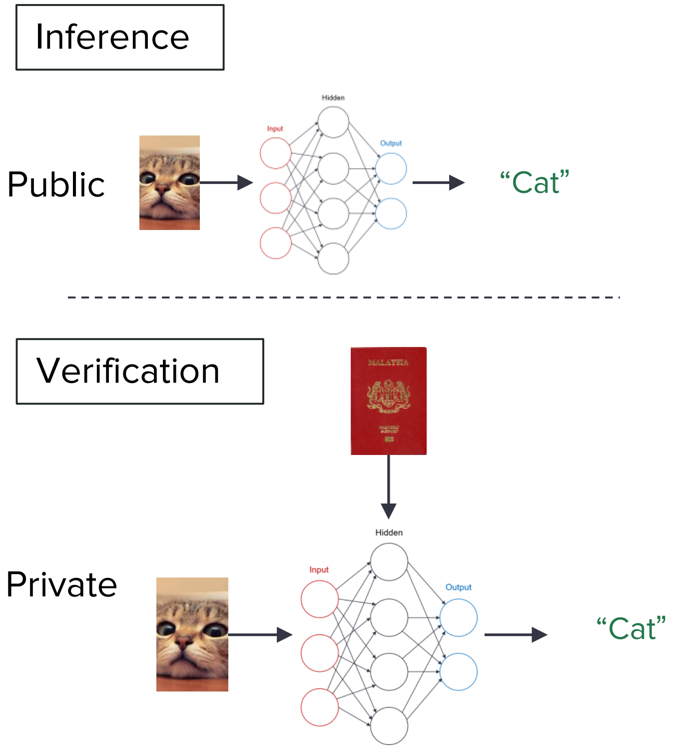
Verification type:

- **White-box**

Disadvantage:

1. **Need to distribute the passports**
2. **Extra inference time**

Ownership Verification with Passports (Scheme 2)



Private passport is embedded but not distributed

Verification type:

- **White-box**

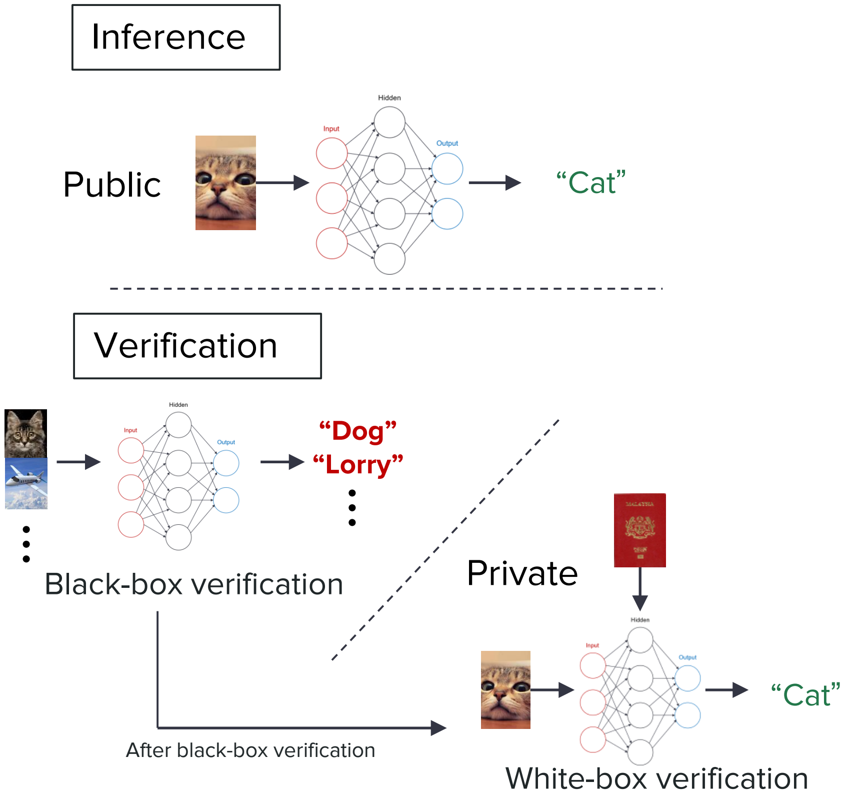
Advantage:

1. **No need to distribute the passports**
2. **No extra inference time**

Disadvantage:

- **Still a white-box verification**

Ownership Verification with Passports (Scheme 3)



Both the private passport and trigger set are embedded but not distributed

Black-box model ownership is verified by **query API calls**

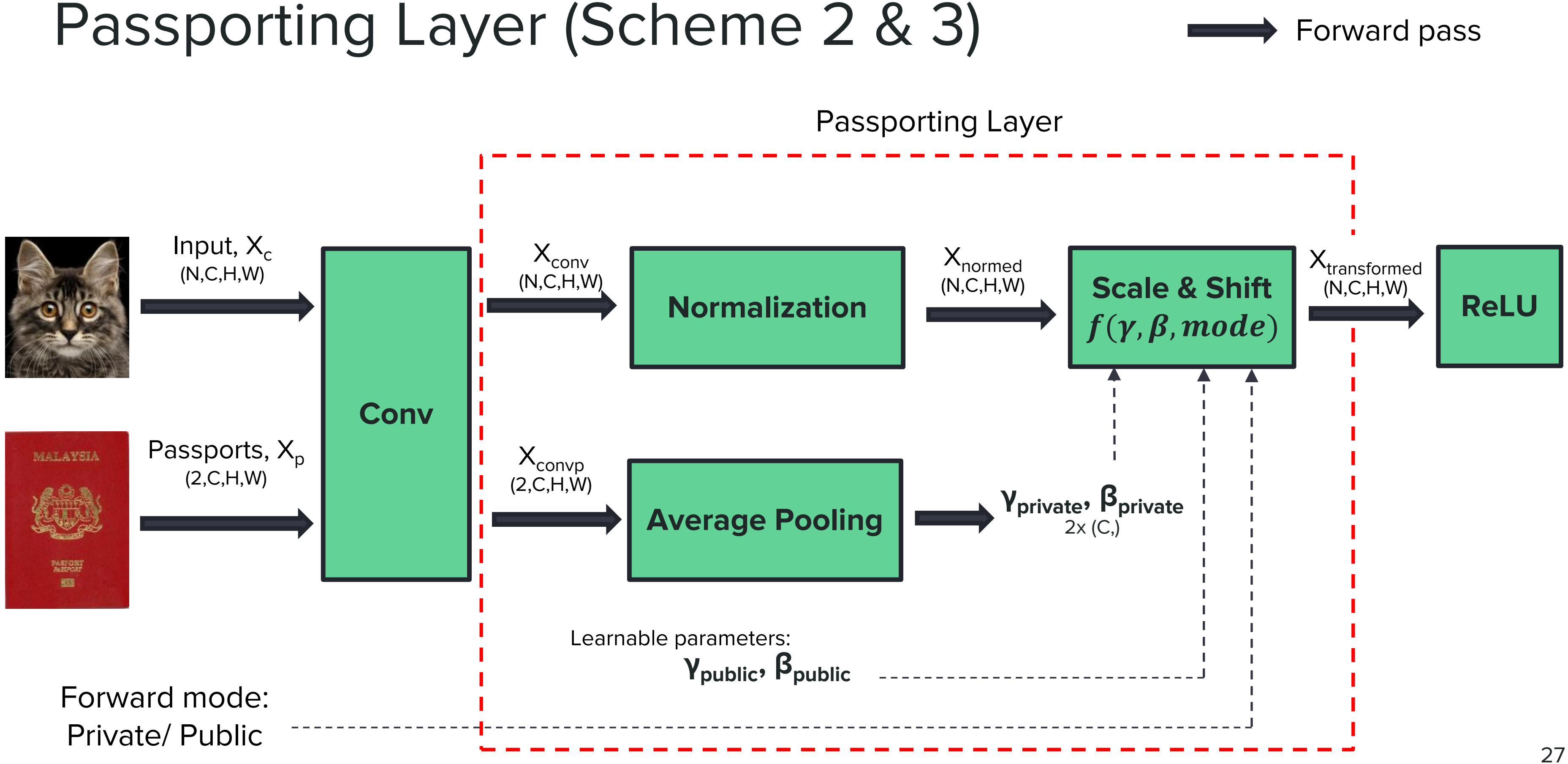
Verification type:

1. **Black-box**
2. **White-box**

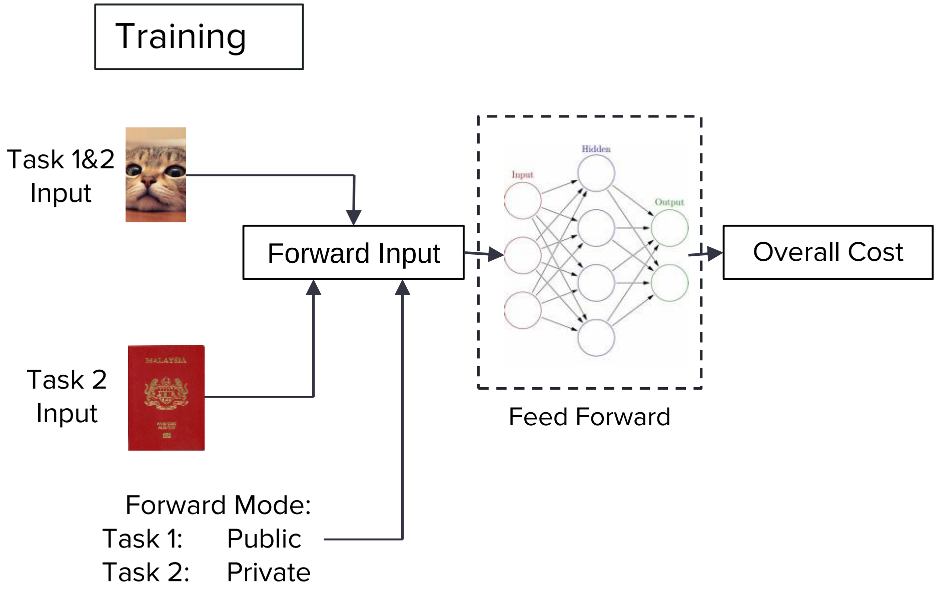
Disadvantages of Scheme 1 & 2 Solved:

1. **No need to distribute the passports**
2. **No extra inference time**
3. **Able to have an initial suspect through black-box verification**

Passporting Layer (Scheme 2 & 3)



Training on Scheme 2 & 3



Multi-task training:

Simultaneously minimizing following cost functions:

Scheme 2

Task 1: Cross Entropy (Public)

Task 2: Cross Entropy + Sign Loss (Private)

Overall Cost: L = L_Task 1 + L_Task 2

Scheme 3

Task 1: Cross Entropy (Public)

Task 2: Cross Entropy + Sign Loss (Private)

Task 3: Trigger-set Embedding Loss (Public + Private)

Overall Cost : L = L_Task 1 + L_Task 2 + L_Task 3

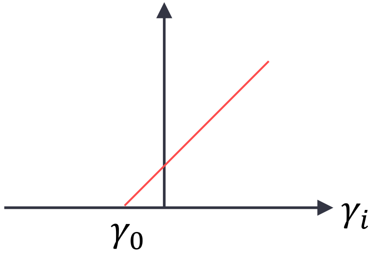
Embedding Binary Signatures by Sign of Scale Factors

- **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**

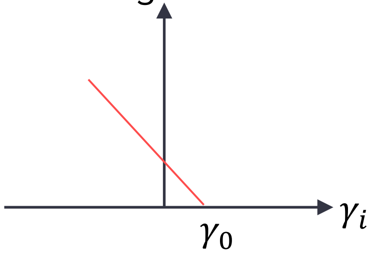
0.1 as default

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

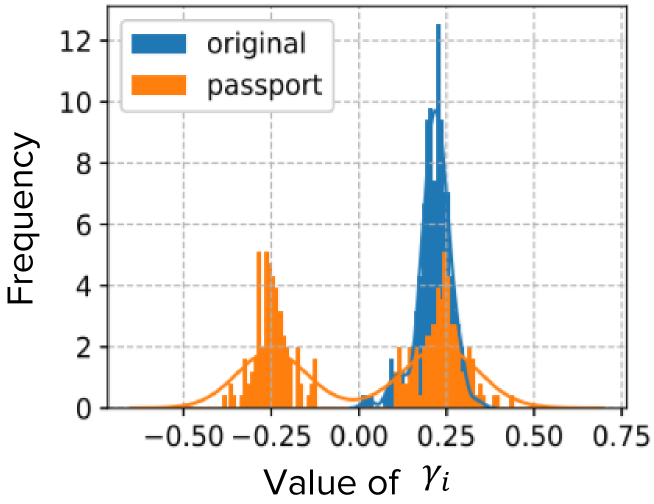
b: [-1 1 -1 1 ...]



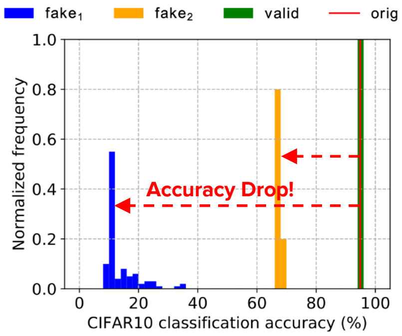
If b is -1 then γ_i need to be negative



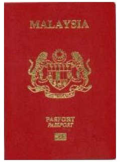
If b is 1 then γ_i need to be positive



Summary of Ambiguity Attacks



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

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**

Summary of Ownership Verification Schemes

	Scheme 1	Scheme 2	Scheme 3
Need to distribute passport	**Yes**	**No**	**No**
Inference time	Up to 10%** more time	**No extra time**	**No extra time**
Training time	Up to 30%** more time	Up to 150%** more time	Up to 150%** more time
Black or **White** box Verification	**White**	**White**	**Black** & **White**

Time increases **are linearly depending on complexity of the network architecture

Take Home message

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

- **Existing** watermarking approaches are **vulnerable to ambiguity attack**

- **Passport-based approach** provided better protection in terms of **robustness against removal attack and ambiguity attack**

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

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.

研讨会官网：
扫描二维码

活动介绍链接：
扫描二维码