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

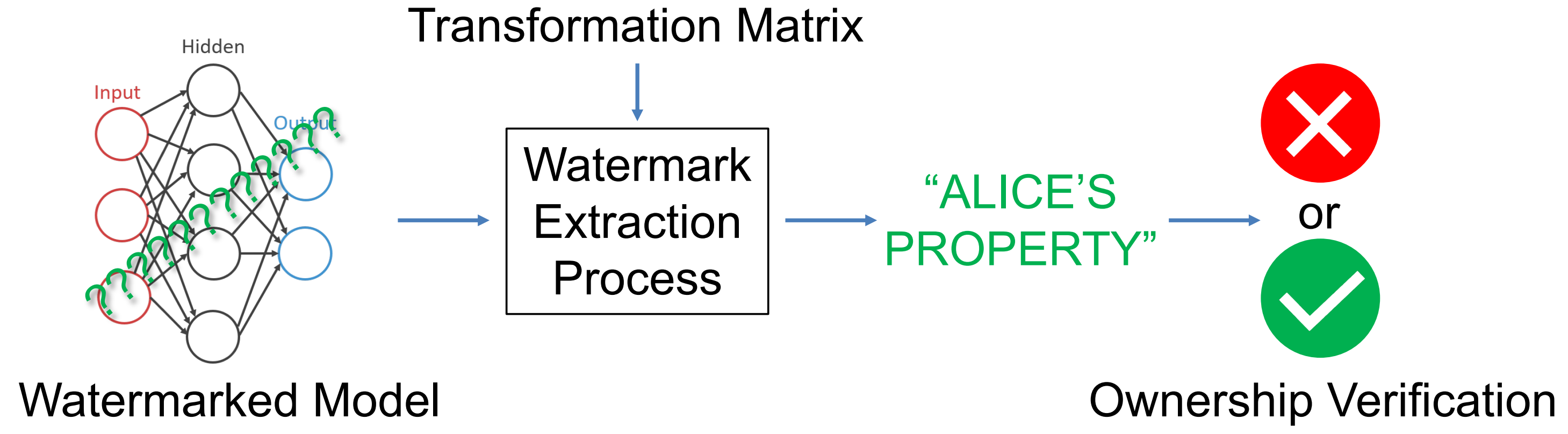
Lixin Fan[1], Kam Woh Ng[1,2], Chee Seng Chan[2]
WeBank AI Lab[1], University of Malaya[2]

2019

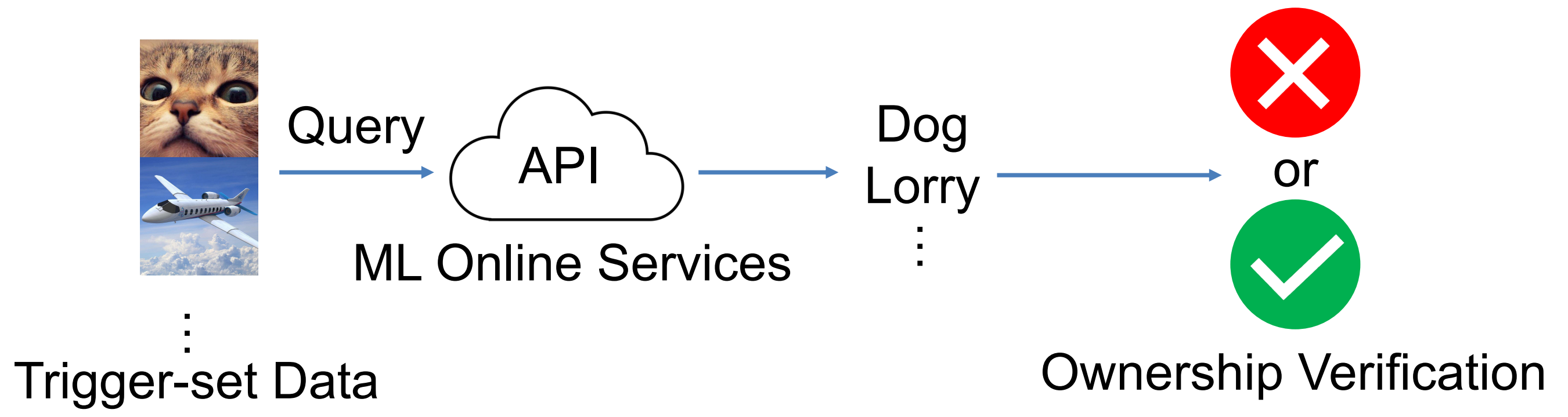
Problem Definition

Conventional DNN Watermarking methods

- **White-box Ownership Verification (Uchida et al. [1])**

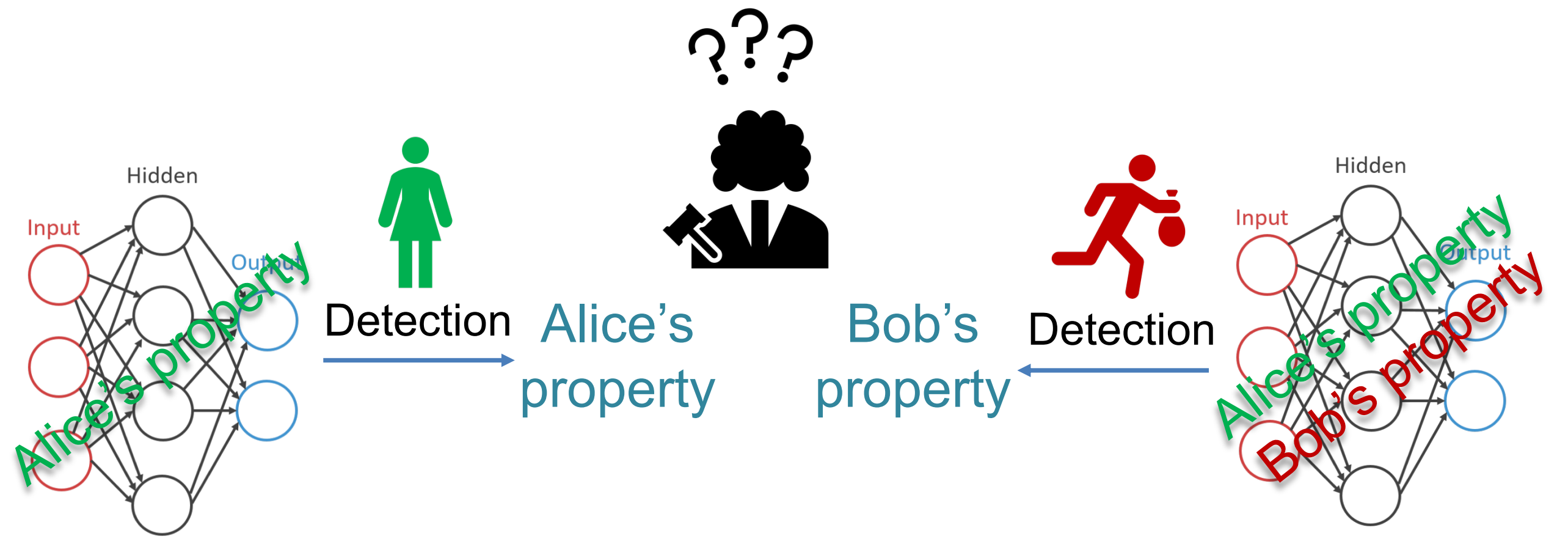


- **Black-box Ownership Verification (Adi et al. [2])**



Problem Statements

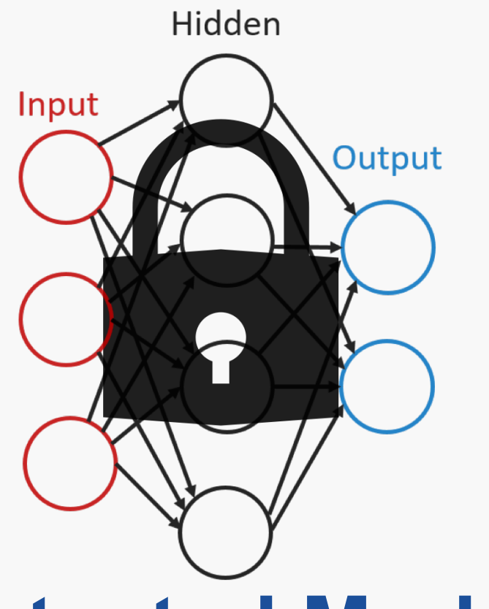
1. Protection on DNN is urgently needed
2. Existing watermarking approaches are vulnerable to ambiguity attack



Watermark Approach	**Real** Watermark	**Fake** Watermark
White-box (Uchida et al. [1])	100% watermark detected	100% watermark detected
Black-box (Adi et al. [2])	100% watermark detected	100% watermark detected

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

Protect your DNN models from theft !

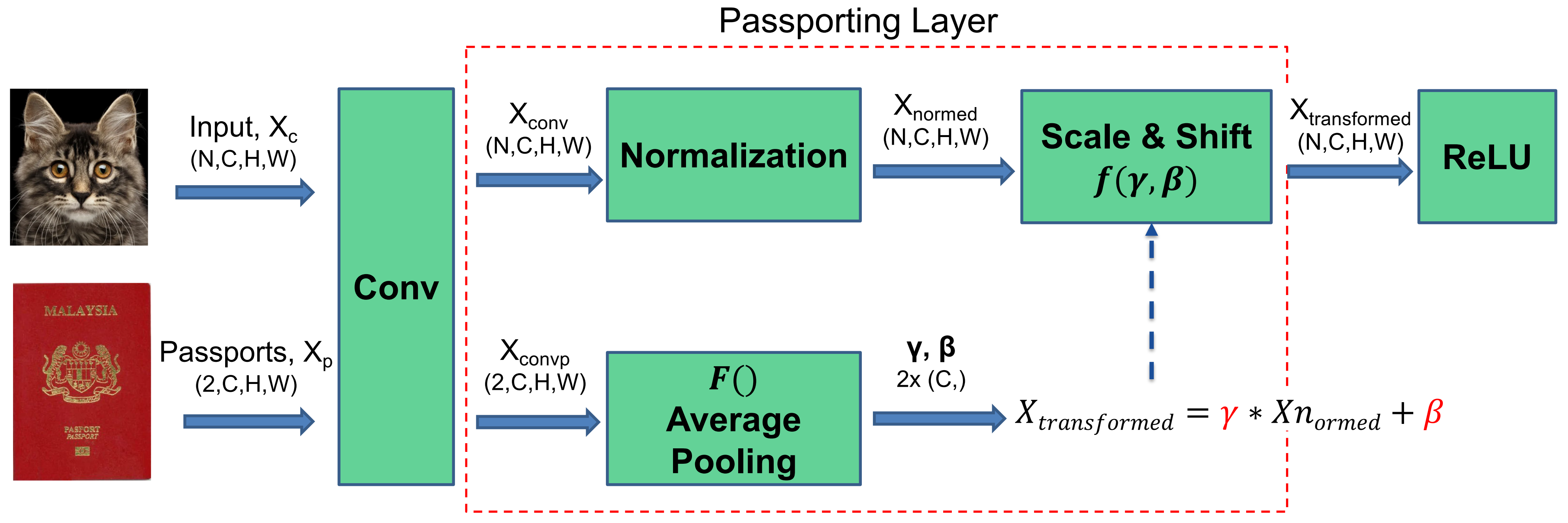


Protected Model

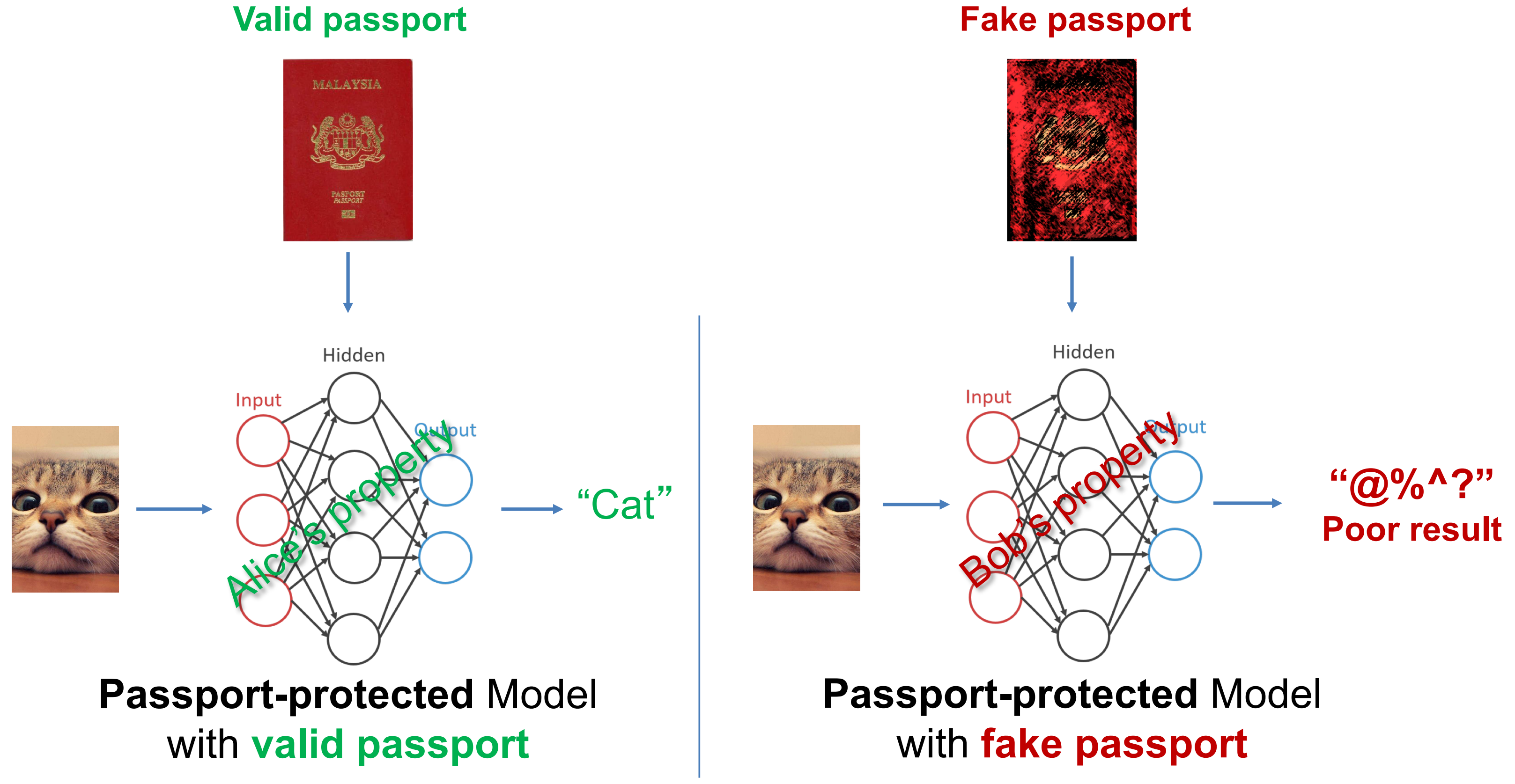
Code & More Details

Our Solution

Passporting Layer



Embedding Passport



Contributions

1. Novel passport-based verification schemes to defeat ambiguity attack
2. One passport-protected DNN model will only have one unique signature
3. Fake passport or modified signature will paralyze the DNN model

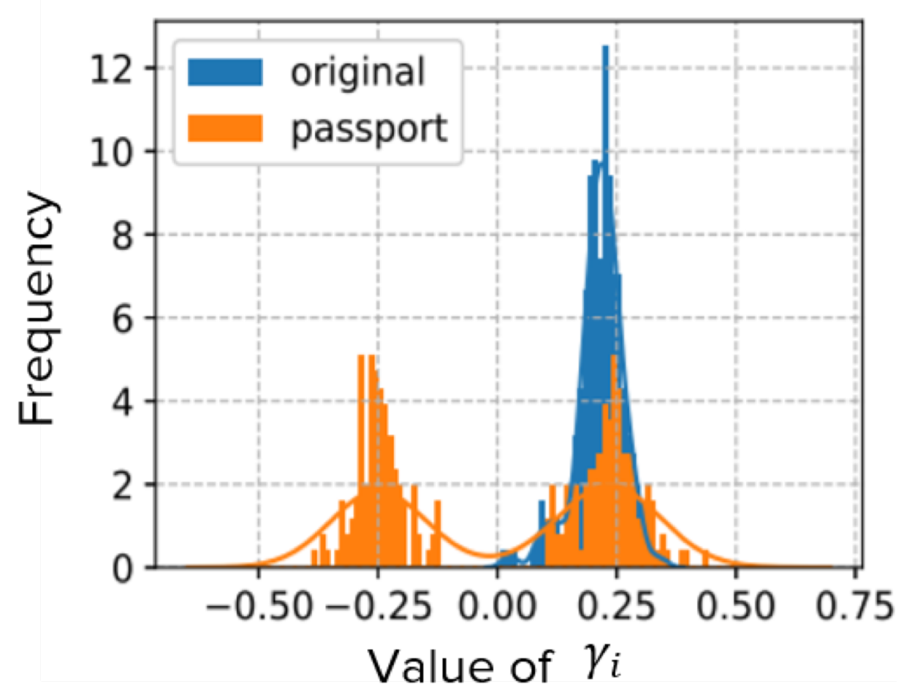
Discussion

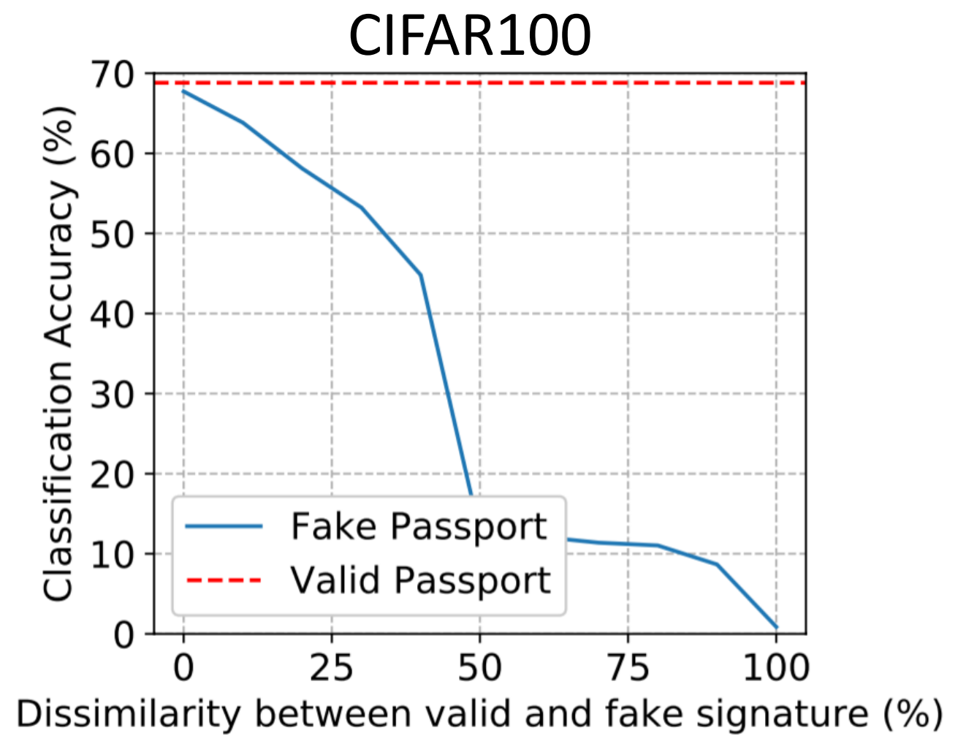
Embedding Binary Signatures into γ of Passporting Layer

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

$$\gamma_0 = 0.1$$
$$b: [-1\ 1\ \ldots]$$

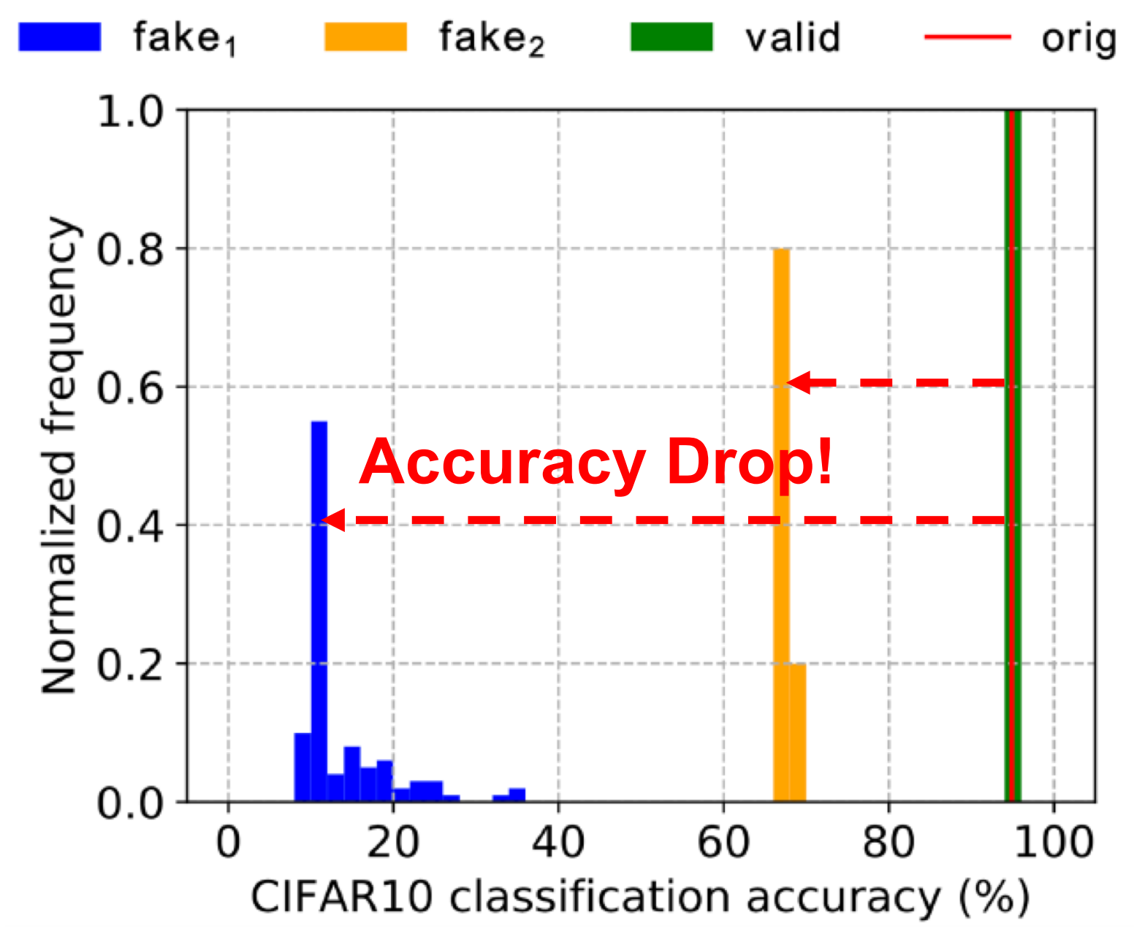
64 channels can embed 8 bytes signature





Experimental Results

Ambiguity attack	Inference Phase	Verification Phase
Fake$_1$ (random passport)	Random guessing	Useless Infringement
Fake$_2$ (reverse-engineered passport)	Performance deteriorated (at best 70% on CIFAR10)	Useless Infringement
Fake$_3$ (copied passport)	Performance Detained Signature Detected	Ownership Verified



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