FedIPR: Ownership Verification for Federated Deep Neural Network Models

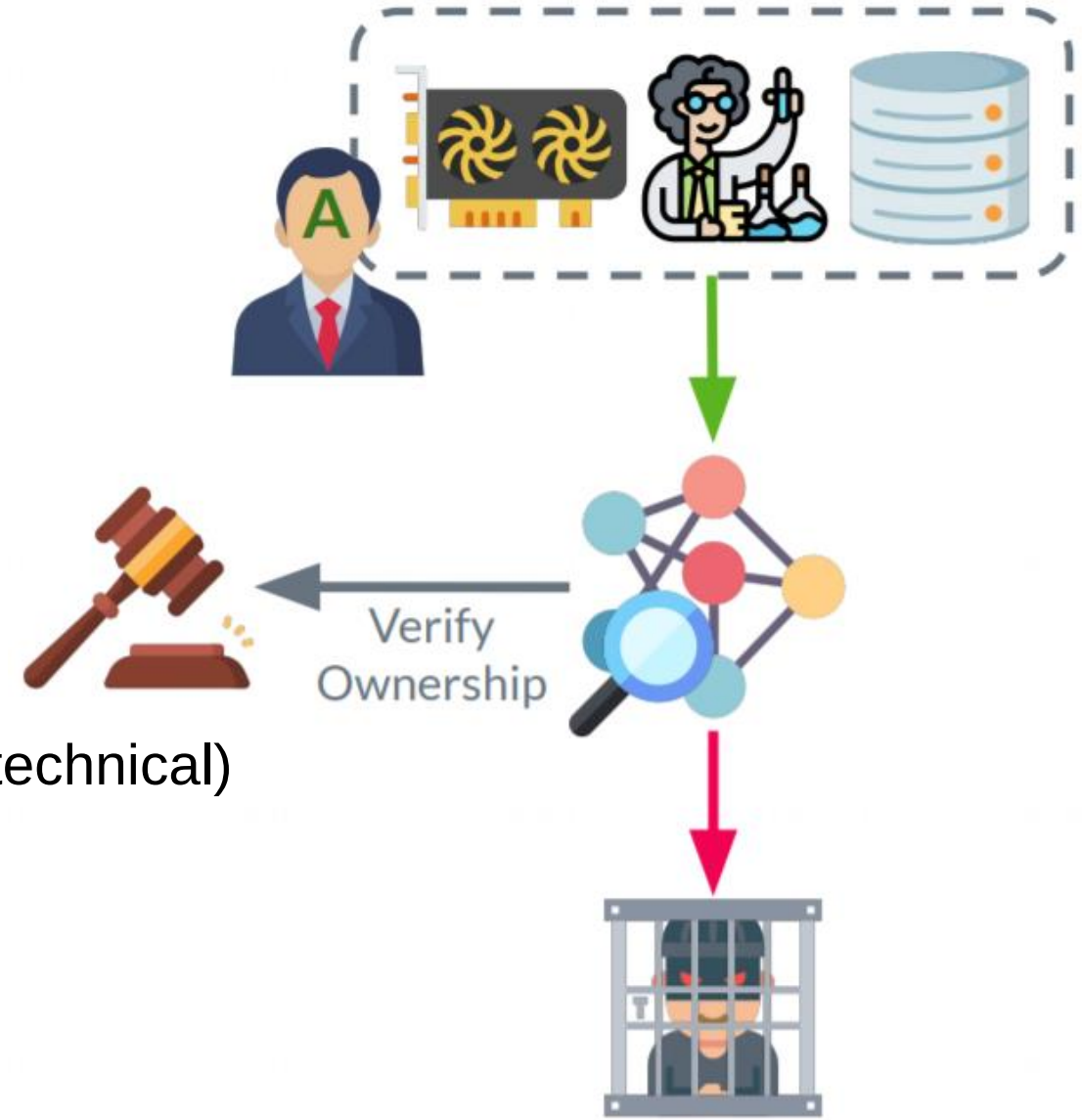
Lixin Fan [1] , Bowen Li [2] , Hanlin Gu [4] , Yan Kang [1] , Jie Li [2] and Qiang Yang [3]

1 AI Group, WeBank Co., Ltd, Shenzhen, China
2 Department of CSE, Shanghai Jiao Tong University, Shanghai, China
3 Department of CSE, Hong Kong University of Science and Technology, Hong Kong, China
4 Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong, China

Background

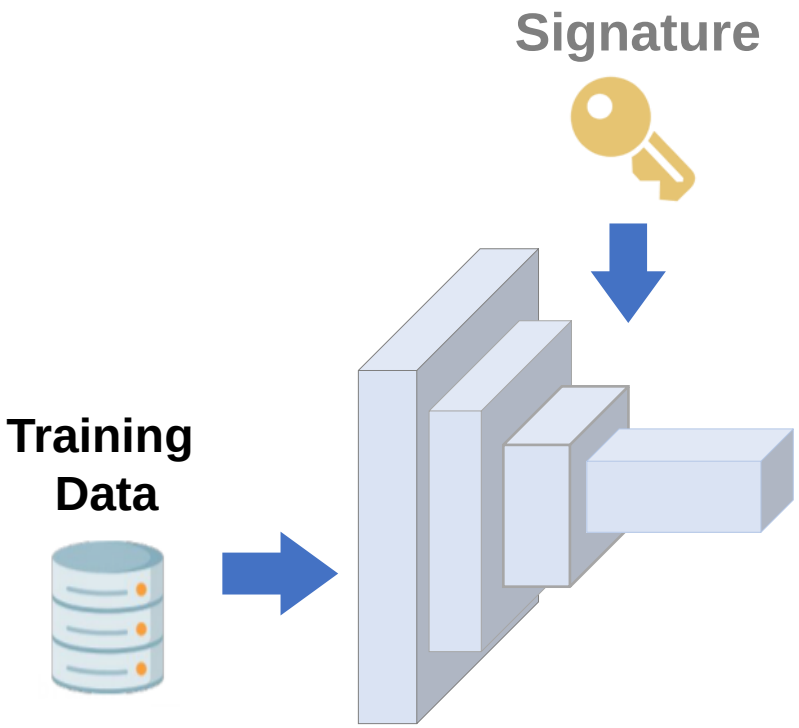
Intellectual Property Right Protection is Necessary:

- Training a DNN is resource intensive
- High business value in trained DNN
- Adversaries may steal the DNN models(Non- technical)
- Verify the ownership of DNN

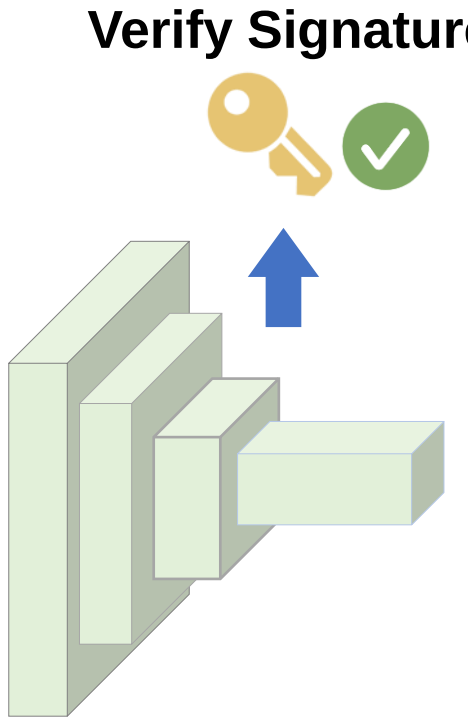


Signature Embedding and Verification

Step 1: Embed distinctive signature when training DNN
Step 2: Detect the prescribed signature from the DNN afterwards



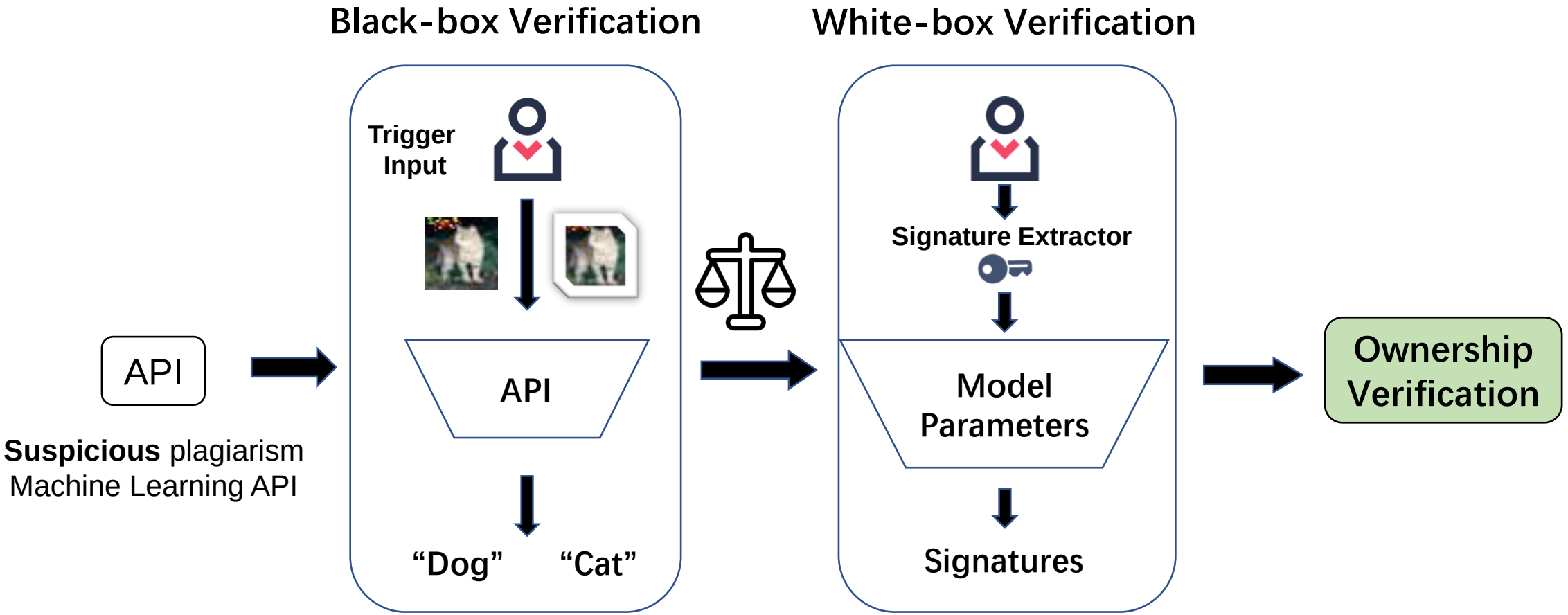
Step1: Embedding while training



Step2: Verification

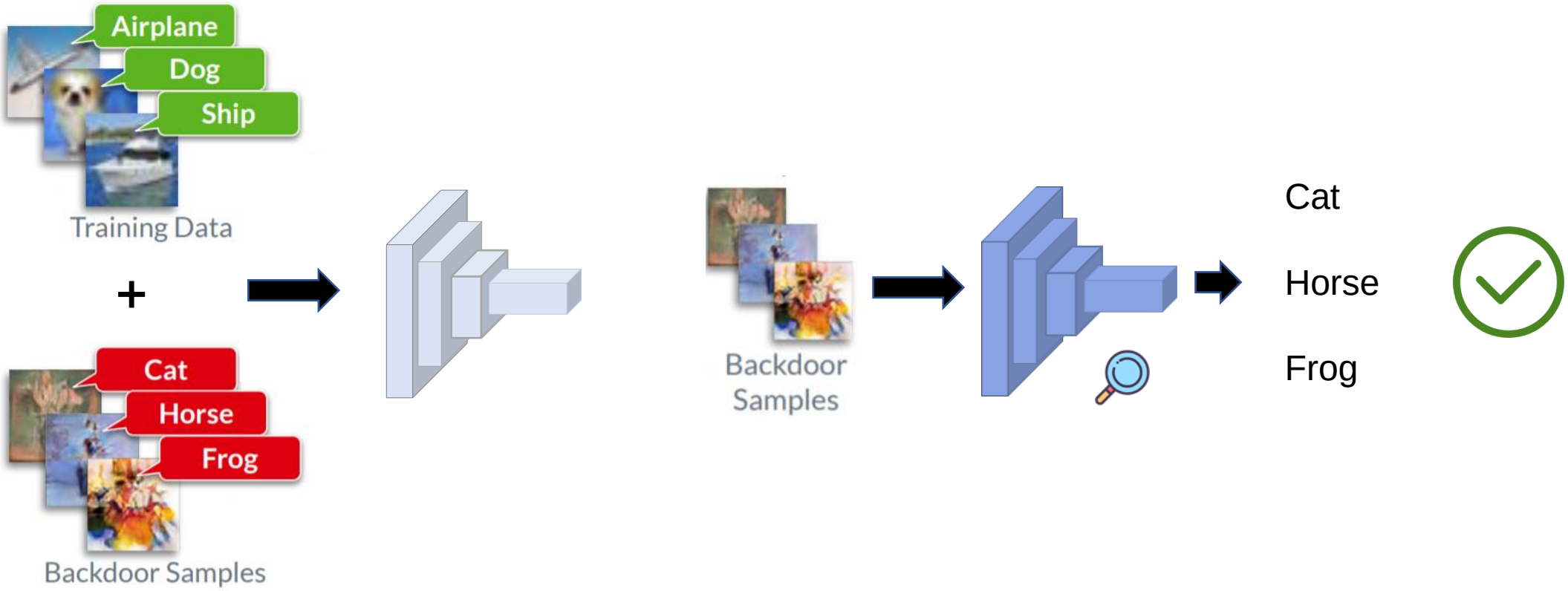
Overview of Verification Process

Black-box : without access to network parameters.
White-box : with access to network parameters.



Previous Work on deep Learning

Black box manner: **Trigger-set based** watermarking proposed by Adi et al.



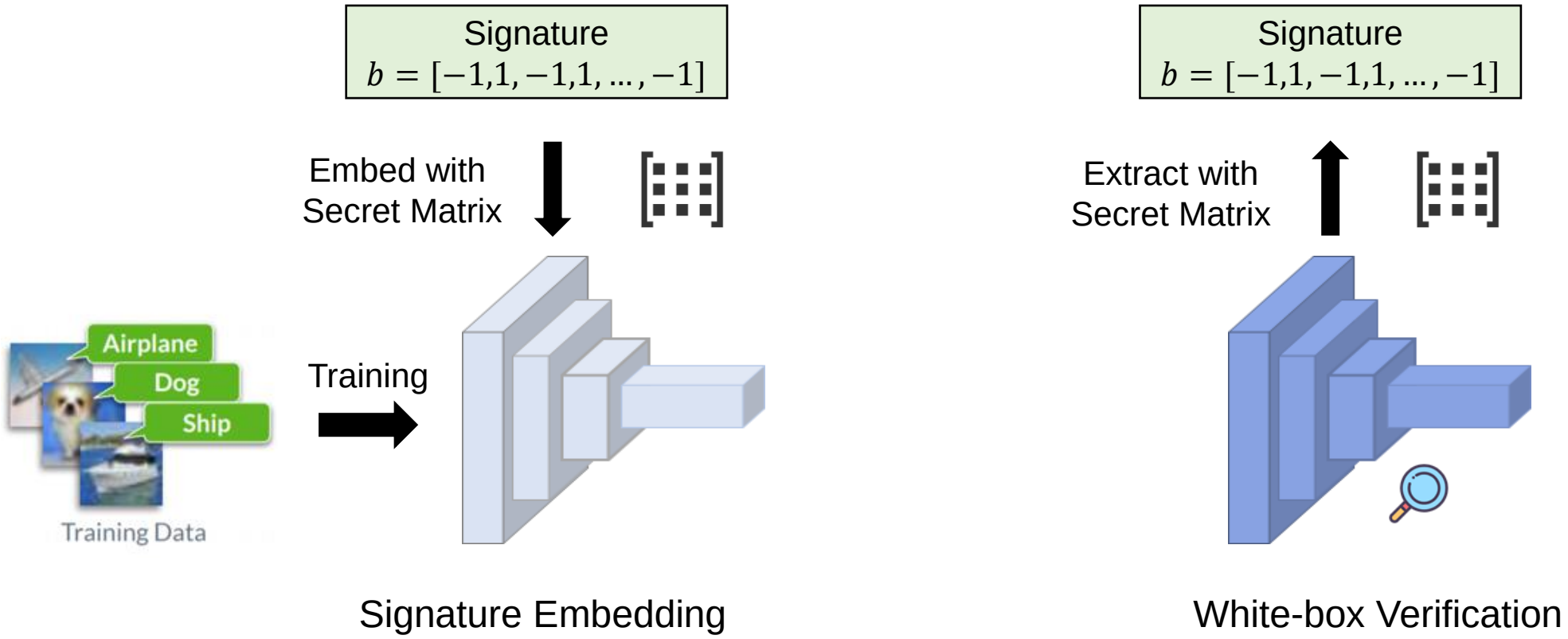
Watermark Embedding

Black-box Verification

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

Previous Work on Deep Learning

White box manner: **Feature based** signature proposed by Uchida et al.

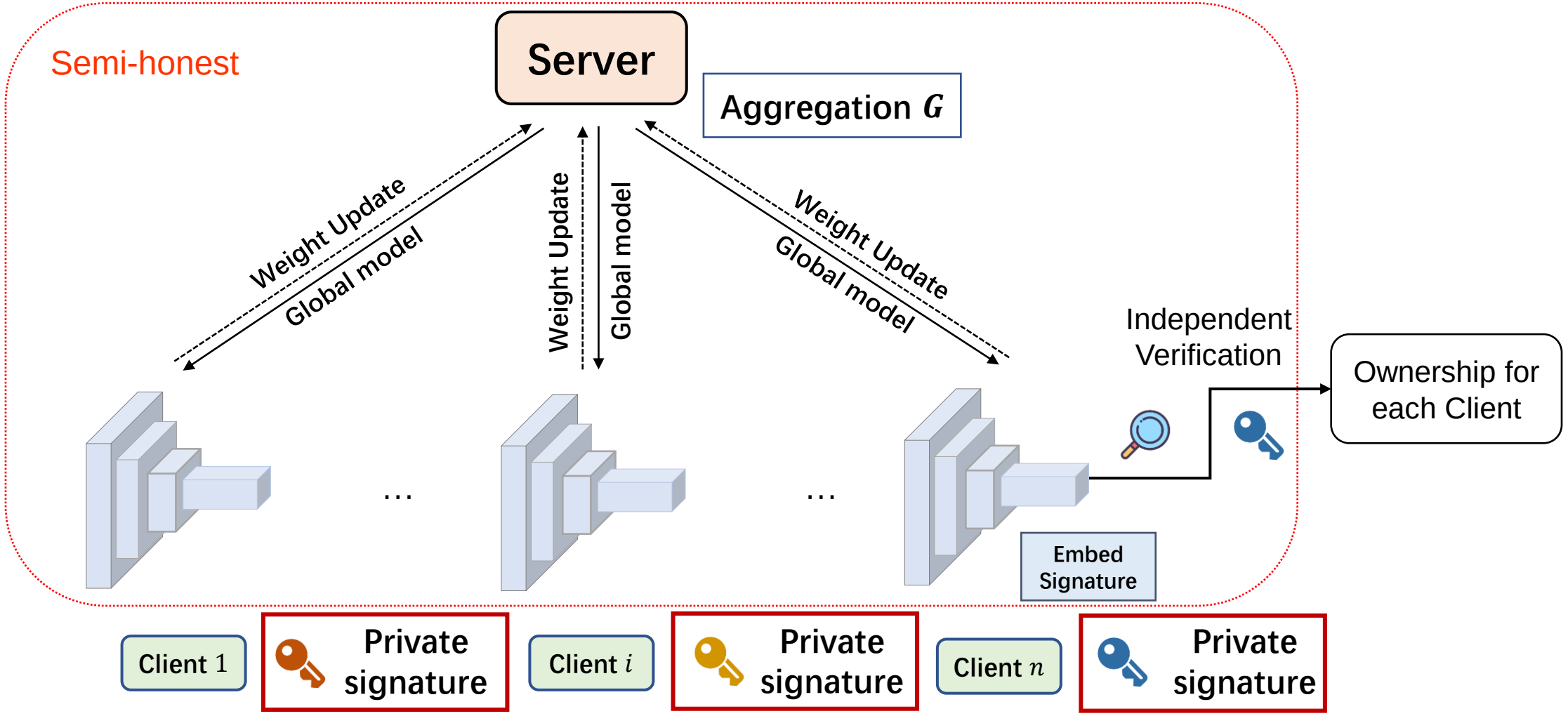


Yusuke Uchida, Yuki Nagai, Shigeyuki Sakazawa, and Shin'ichi Satoh. Embedding watermarks into deep neural networks. In Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, pages 269–277,2017

Goal: Federated Learning and Private Signatures

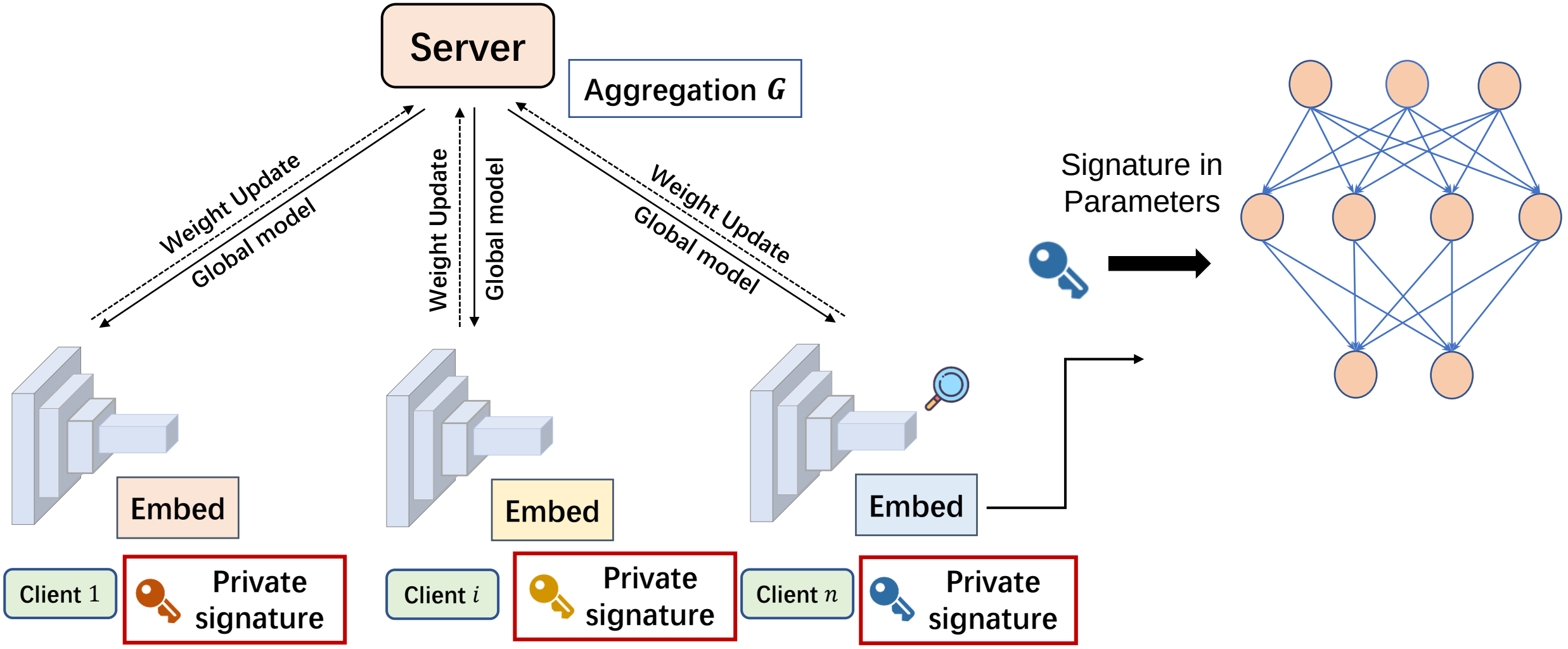
In semi-honest Horizontal Federated learning

How to embed signatures for each client ? Embed signatures for federated common wealth



Ownership Verification in Horizontal Federated Learning (Semi-honest)

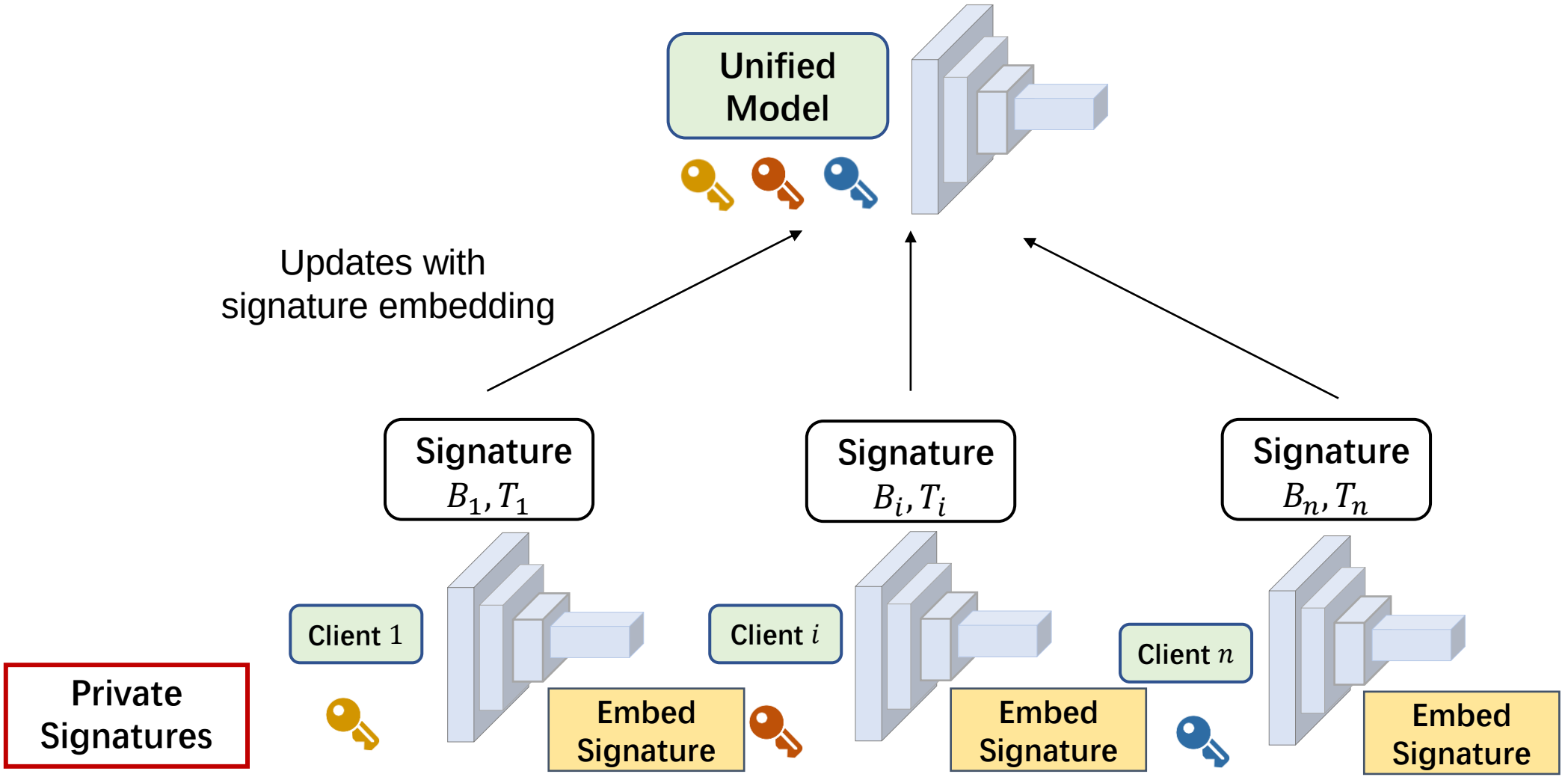
Step1. Embed distinctive Private signature while training



Problem:

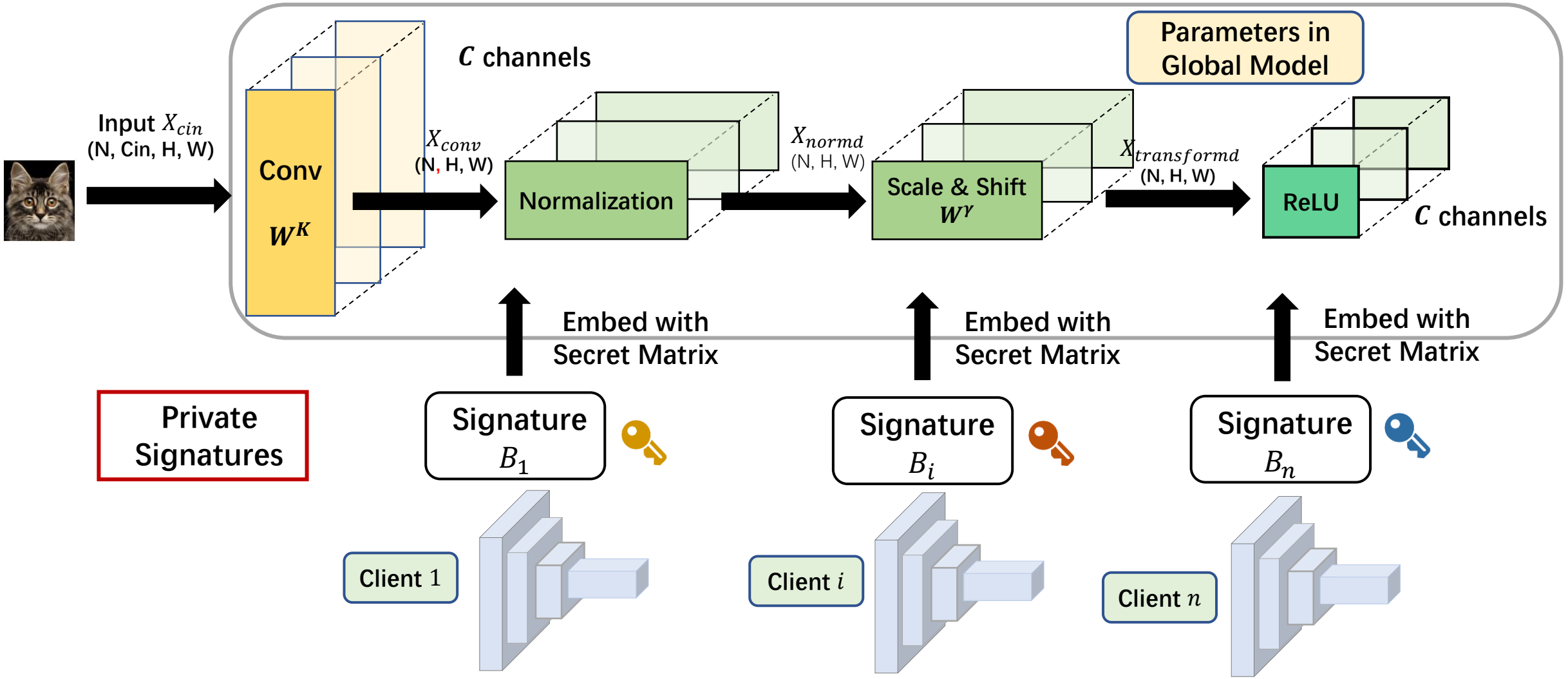
- In a horizontal federated learning system, each client has its own private B_i, T_i,

Signatures are aggregated into one unified model, conflict?



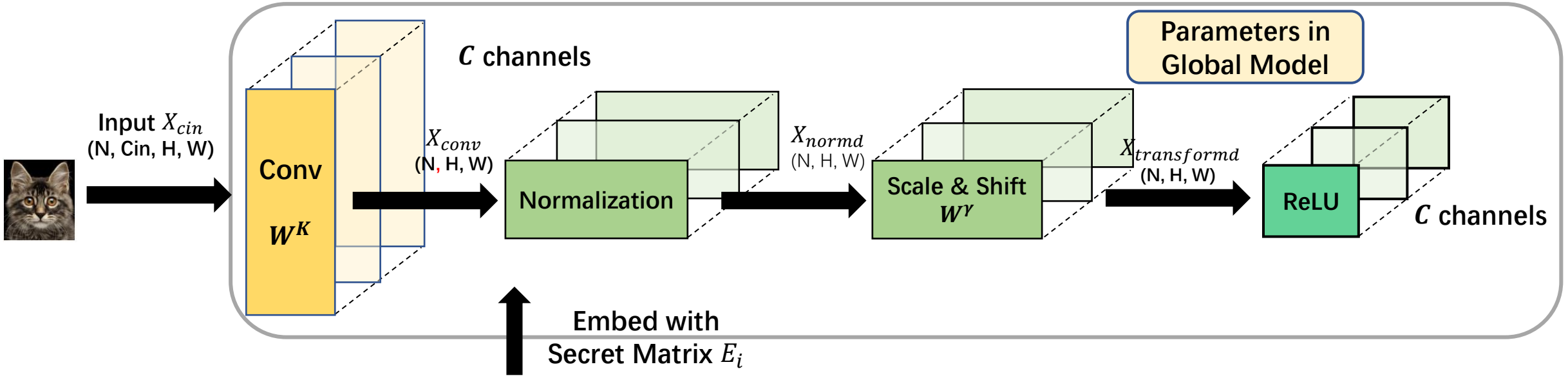
FedIPR: Signatures in Horizontal Federated Learning (Semi-honest)

In a horizontal federated learning system, each client embed its own private B_i, T_i



FedIPR: Signatures in Horizontal Federated Learning (Semi-honest)

In a horizontal federated learning system, each client embed its own private B_i, T_i



$$m = E_i W$$

Sign Loss $L_{sign} = \sum_{i=1}^{n} (\gamma_0 - \gamma_i m_i, 0)$

L_T for T_i trigger-set backdoor

$$L = L_{main} + L_T + L_{sign}$$

Theoretical Proposition

In a n client federated learning system,

when the total bitlength of signature $\sum_{i=1}^{n} l_i$ <= Available channel number in model architecture. The signatures do not conflict.

Definition 2. Let $\mathbf{U}^{M\times KN}$ be matrix combined with $\{\mathbf{E}_1^{M\times N}, \mathbf{E}_2^{M\times N}, \cdots, \mathbf{E}_K^{M\times N}\}$ by column. Let $\tilde{\mathbf{U}}^{M\times KN}$ be matrix combined with $\{(\mathbf{B}_1\mathbf{E}_1)^{M\times N}, (\mathbf{B}_2\mathbf{E}_2)^{M\times N}, \cdots, (\mathbf{B}_N\mathbf{E}_K)^{M\times N}\}$ by column, where $\mathbf{B}_k = (t_{k1}, t_{k2}, \cdots, t_{kN}) \in \{+1, -1\}^N$, is signature of k_{th} client.

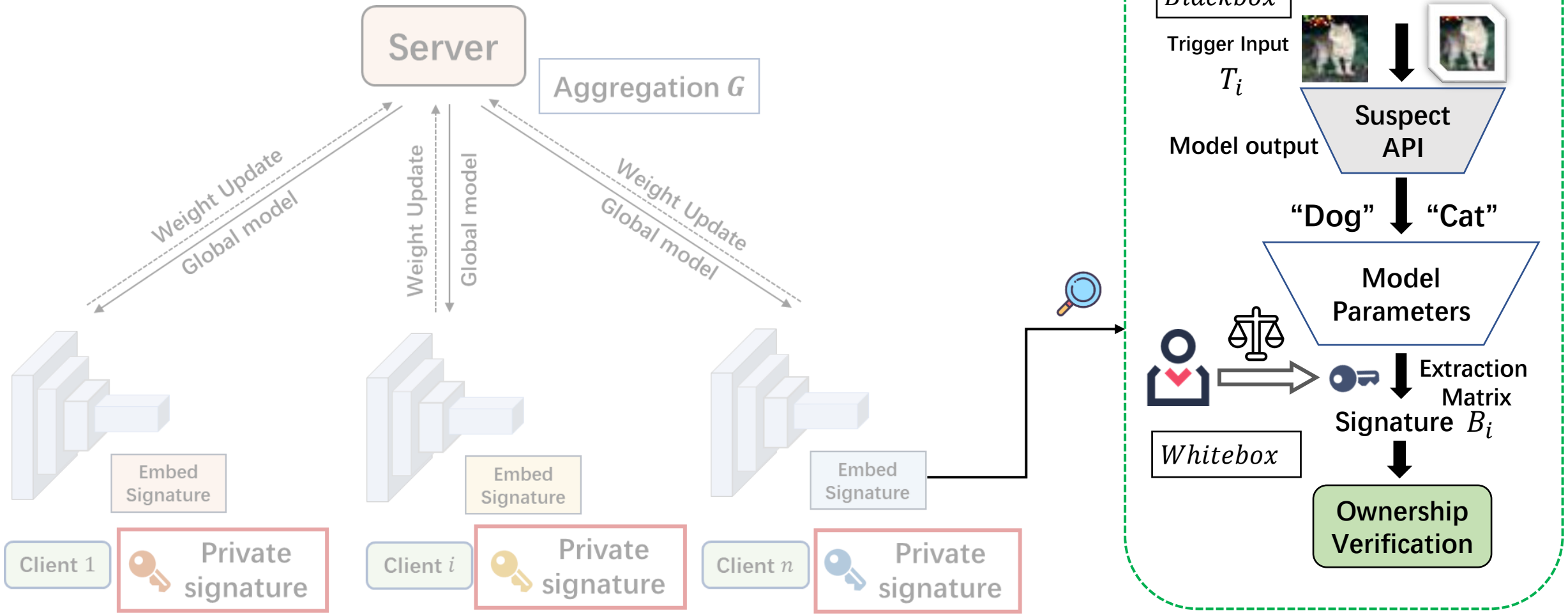
Proposition 1. If $\mathbf{U}$ or $\tilde{\mathbf{U}}$ as defined above satisfy any one of following conditions, then there exists $\mathbf{W}$ such that $\mathbf{W}^T\tilde{\mathbf{U}} \geq 0$.

1. $rank(\mathbf{U}) = KN$,
2. $\exists$ all elements of one row of $\tilde{\mathbf{U}}^{M\times KN}$ are positive,
3. The dot product of any two columns of $\tilde{\mathbf{U}}^{M\times KN}$ are positive.

In addition, when the feature-based sign loss is binary cross-entropy regularization $BCE_{\mathbf{B},\theta}(\mathbf{W^t})$, there exits the common model parameters $\mathbf{W}$ under three conditions such that $\mathbf{WU}$ is less than zero ($\sigma(\mathbf{WU}) < 0.5$) as target signature $\mathbf{B}$ is 0, or larger than zero as target signature $\mathbf{B}$ is 1.

Ownership Verification in Horizontal Federated Learning (Semi-honest)

Step2. Detect the prescribed signature



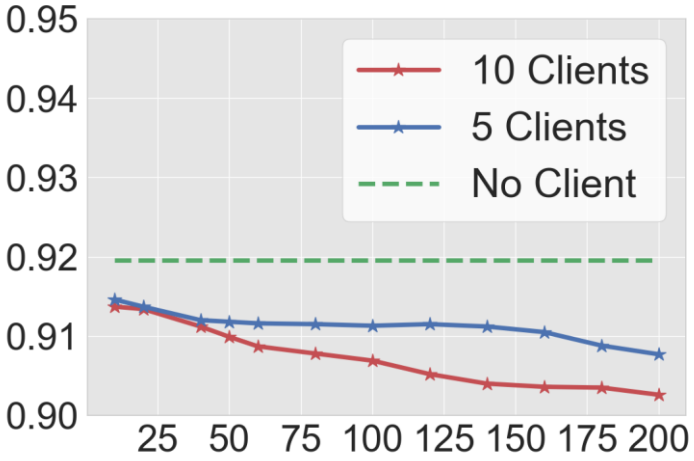
Experiment Results: Fidelity

In a horizontal federated learning system with 20 clients.

Fidelity: Model main task accuracy

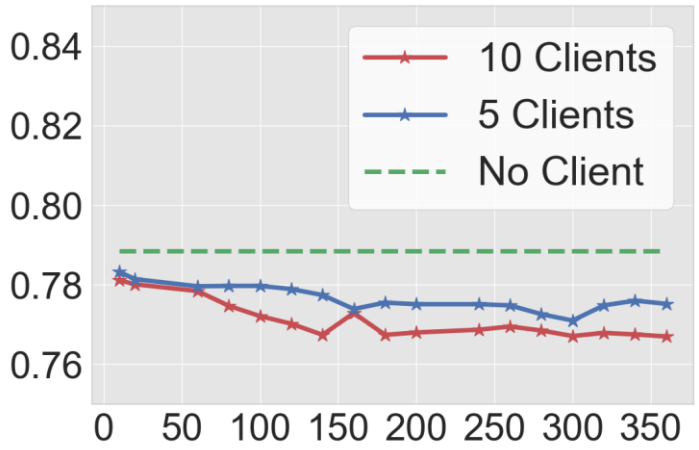
Main task accuracy drop within 2 percent

(1)



Bit Length Per Client

(2)



Bit Length Per Client

Figure (1) and (2), respectively, illustrate CIFAR10 with AlexNet and CIFAR100 with ResNet18 classification accuracy, when 5, 10 clients embed varying bit-lengths signatures.

Experiment Results: Reliability of Signature

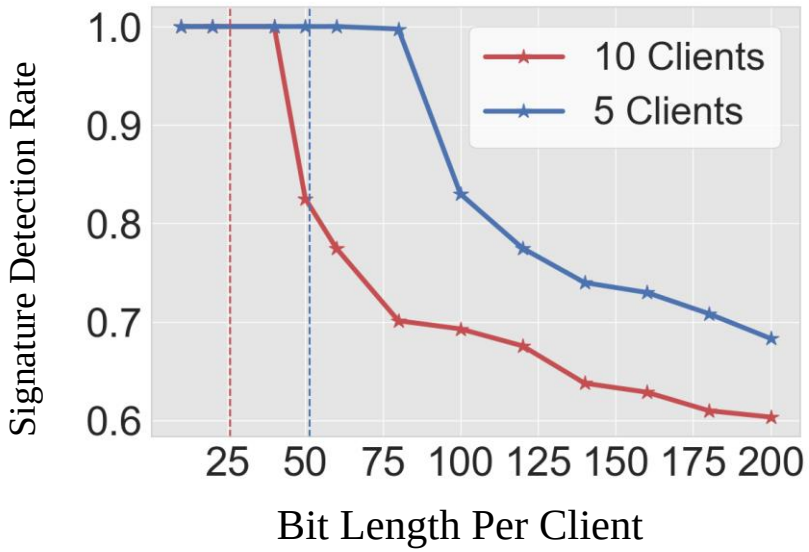
In a horizontal federated learning system with 20 clients.

Reliability: White-box Signature detection rate

The Reliability decades when embedded with feature based signature?

Consistent with theoretical analysis

(1)



(2)

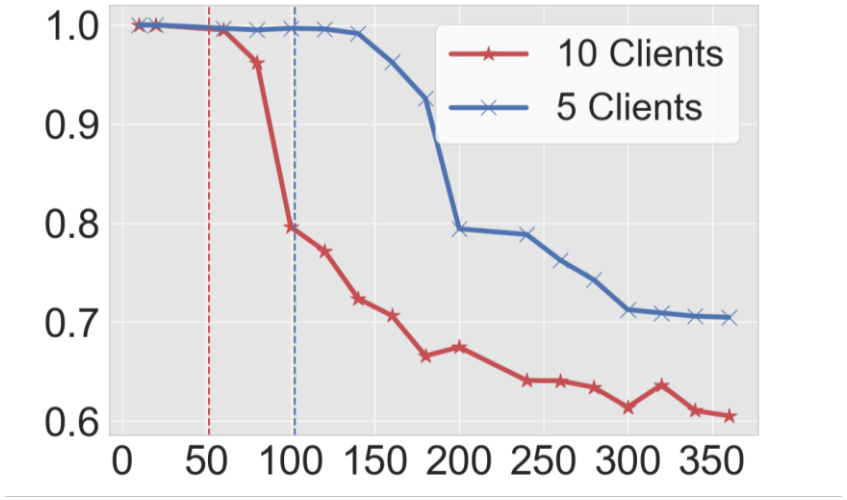


Figure (1) and (2), respectively, illustrate CIFAR10 with AlexNet and CIFAR100 with ResNet18 signature detection rate , when 5, 10 clients embed varying bit-lengths signatures.

Experiment Results: Reliability of Watermark

In a horizontal federated learning system with 20 clients.

Reliability: Black-box Watermark detection rate

Results on defensive methods like Trimmed-Mean, Krum, Bulyan [1, 2, 3] are employed

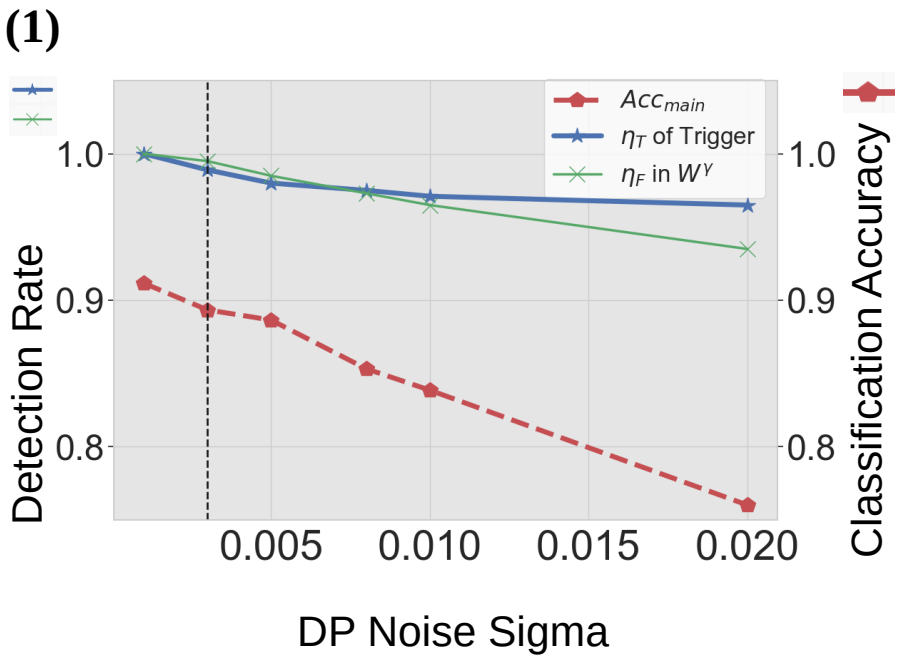
Method	Bulyan	Multi-Krum	Trim-mean	FedAvg
Trigger number per client	80	80	80	80
Detection Rate	68.67%	79.82%	63.25%	98.82%
P(plagiarism)	1-7.21e-78	1-5.24e-35	1-1.74e-47	1-4.02e-30

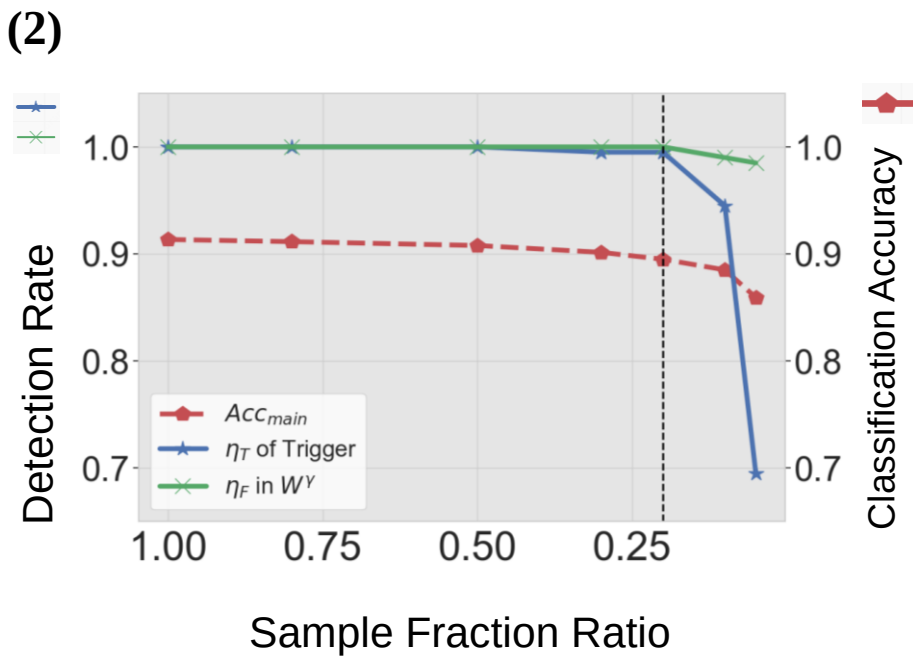
Detection rate twice higher than 1/C is enough to prove plagiarism

[1] P. Blanchard, R. Guerraoui, J. Stainer, et al. Machine learning with adversaries: Byzantine tolerant gradient descent. In Advances in Neural Information Processing Systems, pages 119–129, 2017
[2] E. M. E. Mhamdi, R. Guerraoui, and S. Rouault. The hidden vulnera- bility of distributed learning in byzantium. In ICML. PMLR, 2018.
[3] D. Yin, Y. Chen, K. Ramchandran, and P. L. Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In ICML. PMLR, 2018

Experiment Results: Robustness

Signature Robustness under Federated Learning setting: DP noise, Client sampling





Experiment Results: NLP task

- Federated signature embedding experiments on DistilBert Architecture

Task 1: Sentiment Classification/SST2

Task 2: Text Classification/AG News

Task 3: Natural Language Inference/QNLI

Task 4: Paraphrase/ MRPC

Task/ Dataset	Task 1	Task 2	Task 3	Task 4
Bit-length per client	80	80	80	80
Fidelity in FL	87.59%	89.32%	83.54%	82.62%
Detection Rate	100.00%	100.00%	100.00%	100.00%
P(plagiarism)	100%	100%	100%	100%

Conclusion

- Model IPR Protection is an important demand
- White box /Black box signature provide reliable verification performance
- We evaluate our FedIPR scheme on various computer vision task and model architectures