MarginFinger: Controlling Generated Fingerprint Distance to Classification boundary Using Conditional GANs

Weixing Liu, Shenghua Zhong
College of Computer Science and Software Engineering, Shenzhen University
Guangdong Laboratory of Artificial Intelligence and Digital Economy(SZ)
liuweixing2022@email.szu.edu.cn, csszhong@szu.edu.cn

Motivation
• Training deep neural networks (DNNs) is costly, and attackers can steal models through internal leaks or API access.
• Existing watermarking techniques may lead to a decrease in utility or introduce security risks.
• Recent fingerprinting techniques generate fingerprint samples near the classification boundary to detect pirated models. However, these methods lack distance constraints and are vulnerable to attacks that alter the classification boundary. Some fingerprint samples may change their original labels due to being too close to the classification boundary, leading to their invalidation.

Proposed Method

Figure 1. Principle of our method.

Too far: Both pirated model and irrelevant model show high similarity with the source model’s predicted labels.
Too close: Both pirated model and irrelevant model show low similarity with the source model’s predicted labels.
Just right: Pirated model show high similarity with the source model’s predicted labels.
irrelevant model show low similarity with the source model’s predicted labels.

Step 1: Use a variant of CGAN with margin loss to generate fingerprint samples at a specific distance \(d\) from the source model’s classification boundary.
Step 2: Query the suspect and source model with the generated fingerprint samples.
➢ If the label similarity between them exceeds the threshold \(T\), it’s pirated.

Experimental Results
➢ Dataset: CIFAR-10, Tiny-ImageNet
➢ IP Threats: Model Extraction Probabilities (MEP), Model Extraction Labels (MEL), Model Extraction Adversarial (MEA), Fine-Pruning (FP), Fine-tuning (FT), Transfer Learning (TL)
➢ Baselines: Trigger, IPGuard, CAE, SAC-w, SAC-m

Table 1. The AUC values for baselines facing IP threats in CIFAR-10(left) and Tiny-ImageNet(right), with AVG representing the average AUC. A higher AUC value indicates greater effectiveness.

Conclusion
➢ We propose MarginFinger, a method that utilizes CGAN to generate fingerprint samples, accurately distinguishing between pirated and irrelevant models without the need to train additional models.
➢ We provide a margin loss function to control the distance between the generated data points and the classification boundary, ensuring the robustness and uniqueness of our method.
➢ Extensive results demonstrate that MarginFinger can identify various model theft attacks on different architectures and datasets.
➢ In future work, we aim to enhance the stealthiness of the fingerprint by improving the quality of generated images.