Black Box Watermarking Technology of the AI Model

Yiwei Liu *

College of Computer Science, Chengdu College of University of Electronic Science and Technology of China, Chengdu City, Sichuan Province, 611731, China

* Corresponding Author Email: Liu.yi.wei@outlook.com

Abstract. Today's rapidly developing AI models, black box watermarking technology is becoming an important means of intellectual property protection as it does not require access to the internal structure of the model. The theme of this paper is the black box watermarking technology of artificial intelligence models. First, the paper explains its theory, utilizing the 'memory' of the model to achieve identification through embedding and verification processes. Second, based on the different watermark embedding methods, two mainstream embedding methods are analyzed. In the training embedding stage, the model 'remembers' the mapping relationship, while in the fine-tuning stage, the watermark is embedded with a small number of parameters. Besides, this paper explores key performance indicators including robustness fidelity. Finally, for the current technological development situation, suggestions such as dynamically triggering sample generation are proposed, and future development trends are pointed out. This paper may help the researchers who are working in this field.

Keywords: Watermark, embedding, fin-tuning.

1. Introduction

AI models have become the core asset of enterprises gradually and are widely used in various practical business scenarios. The construction cost of high-performance AI models is high, and their intellectual property protection faces severe challenges, with problems such as easy theft, tampering, or illegal distribution of models. Such behavior not only seriously infringes on the legitimate rights and interests of developers, but also cools the enthusiasm for industrial innovation and disrupts the healthy development of the artificial intelligence market. Thus, the watermark of the AI model is advanced [1].

Watermarking technology embeds concealed ownership information into digital products in an invisible way, achieving property identification and traceability. According to different watermark extraction methods, it can be divided into white box watermark, black box watermark, and unboxed watermark types. Black box watermarking has attracted much attention due to its uniqueness. The core of it is to verify the ownership of the model without accessing the internal structure or parameters of the model, and to complete the verification only through the output results of the model on specific trigger samples. This feature matches the common MaaS pattern in business scenarios, where the model owner only opens the inquiry interface and hides internal details. At heart, it utilizes the characteristics of deep learning models to 'remember' the special correspondence between trigger samples and preset labels. During validation, the trigger samples are input, and the output results are consistent with the preset labels, which proves ownership. Although black box watermarking has shown broad application prospects, the technology is still in a rapid development stage, with different teams proposing different trigger sample designs and watermark embedding strategies, significant differences in evaluation indicators [2], and a lack of unified standards.

Based on the above-mentioned, this paper aims to systematically review black box watermarking technology. By analyzing the technical characteristics and performance of black box watermarking, this article explores the future development trends and research directions of this technology.

2. The Core Theory

The core principle of black box watermarking technology is to utilize the 'memory ability' of specific inputs in AI models, and to achieve ownership verification through preset input-output mapping relationships without relying on internal structural information of the model. Its operation process includes two steps: embedding and verification. In the embedding stage, by adjusting the model training process or fine-tuning the already trained model, the model produces stable preset outputs for the designed trigger samples. This embedding requires strict control over the impact on the original task performance of the model. In order to ensure the value of the model, it is usually required that the decrease in the accuracy rate does not exceed 5%. In the verification phase, only trigger samples need to be input to the model under test. If the output results match the preset rules to a certain threshold value, ownership can be proven.

3. Mainstream Embedding Methods

From a technical perspective, black box watermarking technology can be classified according to watermark embedding strategies, which can be specifically divided into training stage embedding and fine-tuning stage embedding.

3.1. Embedding During the Training Phase

The training embedding stage involves mixing trigger samples into the training set during the early stages of model training, allowing the model to 'remember' the mapping relationship between trigger samples and preset labels while learning the patterns of the original data. The backdoor watermarking scheme proposed by Adi et al. for model copyright authentication uses labeled images (trigger samples) and normal images as training data during image classification model training. The model learns to classify different images and outputs preset labels for trigger samples to achieve watermark embedding [3].

The merit of this method is strong watermark stability, deep fitting between trigger samples and models, and good resistance to fine-tuning attacks. However, the limitation lies in the need to fully participate in model training, which cannot be applied to deployed mature models, and a large number of trigger samples may interfere with the distribution of training data, affecting the accuracy of the original task of the model.

3.2. Fine-tuning Stage Embedding

Fine-tuning refers to the technique of retraining deep layers related to the target to adapt to new classification tasks while retaining some layer weights on the basis of pre-trained convolutional neural networks [4]. In this stage of embedding, a small number of trigger samples are used to fine-tune parameters for the trained model, usually only modifying the top-level or output layer parameters, in order to achieve watermark embedding with minimal cost. The fine-tuning model decision boundary proposed is used to represent watermark information. Gathering two types of samples near the decision boundary, real adversarial samples (misclassified by the original model) and incorrect adversarial samples (still correctly predicted by the original model), train to classify real adversarial samples into real tags, and adjust the decision boundary to give the model a unique decision boundary [5].

The kernel benefit of this strategy is its wide applicability, which can be used for closed-source or third-party pre-trained models without the need to obtain the whole training data. However, the watermark strength is limited by the fine-tuning amplitude. If the fine-tuning amplitude is too small, it is easy to be overwritten by subsequent model updates. If the amplitude is too large, it may significantly reduce the accuracy of the original task.

3.3. Comparative Analysis

As shown in Table 1, a comparison of two methods is presented. During the training embedding phase, it is necessary to generate trigger samples that are highly consistent with the original data distribution through generative adversarial networks, reducing interference with the accuracy of the model's original tasks, and dynamically adjusting the number and strength of trigger samples to adapt to the training process. The fine-tuning embedding stage should use reinforcement learning to optimize the fine-tuning amplitude [6] and balance the watermark strength and model performance, while combining knowledge distillation technology [7] to enhance watermark stability and improve adaptability to different model architectures.

Table 1. Method comparison

Methods	Technical Pathway	Merit	Weakness
Embedding during the training phase	Mixing trigger samples into the initial training set	The watermark model is more stable	Only applicable to new models, unable to adapt to deployed models
Fine-tuning stage embedding	Fine-tune the training model parameters with a small number of trigger samples (focusing on the top/output layer)	Wide applicability, supporting closed-source models	overwritten by subsequent updates

Digital watermarking technology will deeply integrate artificial intelligence, blockchain and other technologies to achieve adaptive embedding, trustworthy traceability and high security protection. The application scenarios are expanding from traditional media to emerging fields such as virtual reality and AIGC, addressing copyright and traceability issues for various types of content. In the future, transparent and open methods should be adopted for watermark verification, and attempts can be made to combine watermark technology with cryptographic methods such as multi-party secure computation and zero knowledge proof [8]. At the same time, technical standards and regulations will gradually improve, enhance cross-system compatibility to promote large-scale applications, and the anti-attack ability, concealment, and computational efficiency of watermarks will continue to be optimized to better adapt to real-time and complex scene requirements.

4. Key Performance Indicators

Robustness refers to the ability of a digital watermark to be accurately detected, extracted, or recognized even after undergoing various intentional or unintentional signal processing, attacks, or tampering. For example, in copyright protection scenarios, if the watermark is easily removed or destroyed, it cannot prove the ownership of the work. As the number of layers embedded with watermarks increases and the number of training epochs increases, the watermark becomes more robust [9,10]. Security refers to the fact that the watermark itself cannot easily be faked. And validity is the 'ultimate test' of robustness, safety, and other indicators. If the watermark is easily damaged by conventional processing (poor robustness) or susceptible to malicious cracking (poor security), then it cannot achieve the preset function and cannot be considered effective. The concept of "high robustness" or "high security" that deviates from specific application scenarios may have no practical significance. In addition, watermark capacity only refers to how much information is carried in the watermark, and the capacity issue needs to be considered in the design process.

5. Conclusion

Overall, this paper focuses on the embedding strategy of black box watermarking, analyzing the technical characteristics of the training stage embedding, fine-tuning stage embedding, and two types of methods and their advantages and disadvantages. During the training phase, the embedding stability is strong, but the applicability is narrow. During the fine-tuning phase, the embedding

flexibility is high, but the anti-attack ability is weak. In the output layer, the additional embedding fidelity is good but the security is insufficient. Future research needs to make breakthroughs in dynamic triggering sample design, multimodal fusion embedding, distributed security verification, and other directions, while promoting the construction of industry standards and legal systems, in order to achieve a leap from laboratory research to industrial implementation of black box watermarking technology and provide reliable support for intellectual property protection of artificial intelligence models.

References

[1] Uchida Y, Nagai Y, Sakazawa S, et al. Embedding watermarks into deep neural networks. Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval. New York: ACM, 2017: 269 - 277.

[2] Xie Chenqi, Zhang Baowen, Yi Ping. Research on watermarking of artificial intelligence models. Computer Science, 2021, 48 (07): 9 - 16.

[3] Adi Y, Baum C, Cisse M, et al. Turning your weakness into a strength: Watermarking deep neural networks by backdooring. Proceedings of the 27th USENIX Security Symposium. 2018: 1615 - 1631.

[4] Cetinic E, Lipic T, Grgic S. Fine-tuning convolutional neural networks for fine art classification. Expert Systems with Applications, 2018, 114: 107 - 118.

[5] Wu Hanzhou, Zhang Jie, Li Yue, et al. Overview of artificial intelligence model watermarking. Journal of Image and Graphics, 2023, 28 (6): 1792 - 1810.

[6] Wołczyk M, Cupiał B, Ostaszewski M, Bortkiewicz M, Zając M, Pascanu R, et al. Fine-tuning reinforcement learning models is secretly a forgetting mitigation problem. arXiv preprint arXiv:2402.02868, 2024.

[7] Papernot N, McDaniel P, Wu X, et al. Distillation as a defense to adversarial perturbations against deep neural networks. Proceedings of 2016 IEEE Symposium on Security and Privacy. 2016: 582 - 597.

[8] Feng Shuai, Deng Lunzhi. Identity-based data auditing scheme with privacy protection. Journal of Guizhou Normal University (Natural Sciences), 2023, 41 (2): 105 − 112.

[9] Sarker J, Turzo AK, Bosu A. A benchmark study of the contemporary toxicity detectors on software engineering interactions. In 2020 27th Asia-Pacific Software Engineering Conference (APSEC). IEEE, 2020: 218 - 227.

[10] Feng Le, Zhu Renjie, Wu Hanzhou, et al. Overview of neural network watermarking. Journal of Applied Science, 2021, 39 (06): 881 - 892.