Enhancing Probabilistic Watermark Robustness in Neural Networks via Bilinear Pooling and Attention-Driven Trigger Generation

Akash Adak
Department of Information Technology,
Indian Institute of Information Technology Allahabad
Allahabad, India
iit2022215@iiita.ac.in

Harsh Ramasurat Yadav
Department of Information Technology,
Indian Institute of Information Technology Allahabad
Allahabad, India
iit2022184@iiita.ac.in

***Abstract*—In this paper, we propose enhancements to the trigger set generation process for probabilistically robust watermarking of neural networks. Building upon the PROWN framework, we introduce bilinear pooling and attention mechanisms during the creation of trigger sets to achieve more complex and transferable watermark patterns. Bilinear pooling captures higher-order feature interactions between sample pairs, while attention-driven feature fusion allows the model to focus on semantically important regions during trigger generation. Our method retains the original PROWN model architecture, verification pipeline, and evaluation setup, ensuring direct comparability. Extensive experiments conducted on CIFAR-10 and CIFAR-100 datasets demonstrate that both bilinear pooling and attention-based trigger generation significantly outperform the original trigger construction in terms of trigger set transferability, robustness against model stealing attacks, and ownership verification accuracy. The proposed modifications introduce minimal computational overhead and can be seamlessly integrated into existing probabilistic watermarking frameworks, providing a simple yet highly effective enhancement for intellectual property protection in machine learning models.**

I. Introduction

A. Overview

The rapid adoption of deep learning models across various domains such as computer vision, natural language processing, and healthcare has raised critical concerns regarding the protection of intellectual property (IP) embedded in trained models. As the cost of data collection, model development, and computational resources increases, ensuring ownership rights over deployed neural networks becomes essential. Watermarking, the process of embedding a hidden signature into a model, has emerged as a promising solution for IP protection. Particularly, trigger set-based watermarking techniques—where a model is trained to produce specific outputs on specially crafted inputs—have gained popularity for black-box ownership verification. Among these, probabilistic approaches such as PROWN (Probabilistically Robust Watermarking of Neural Networks) have shown significant robustness against model stealing and distillation attacks.

B. Motivation

While trigger set-based methods like PROWN offer improved resilience against functionality stealing, a key limitation remains in the simplicity of trigger generation. Standard methods often rely on random convex combinations of inputs, which may not optimally exploit complex feature relationships within data. This limits the watermark's transferability and detectability under adversarial model modifications. Motivated by this observation, we explore the potential of higher-order feature interactions and attention-driven mechanisms to generate richer and more robust trigger sets. By integrating bilinear pooling and attention into the trigger generation process, we aim to create trigger sets that encode deeper semantic relationships between samples, thereby enhancing the probability of transferability to stolen models without sacrificing model performance.

C. Significance and Contributions

In this work, we extend the PROWN framework by introducing two major enhancements during the trigger set generation phase:

- **Bilinear Pooling for Feature Interaction:** We apply bilinear pooling to capture pairwise interactions between feature channels of paired samples, allowing the trigger samples to represent more complex and discriminative patterns.
- **Attention Mechanisms for Semantic Focus:** We leverage an attention-based module to emphasize salient regions within paired samples during trigger generation, ensuring that the model focuses on semantically important features while learning the watermark.

Our enhancements do not require retraining the model architecture or modifying the proxy model verification pipeline, thus maintaining the lightweight and general-purpose nature of PROWN. Through extensive experiments on CIFAR-10 and CIFAR-100 benchmarks, we demonstrate that our approach significantly improves trigger set robustness and watermark verification accuracy compared to baseline methods.

II. Problem Statement

The increasing value of deep learning models has made them prime targets for intellectual property theft through model stealing attacks, where an adversary replicates a model's functionality without access to its original architecture or training data. Watermarking techniques, particularly trigger set-based methods, offer a practical defense by embedding secret patterns into the model's behavior. The owner can later verify ownership by querying the suspect model with a pre-constructed trigger set and checking for specific responses.

However, a major challenge persists: **existing trigger set generation techniques are often based on random data mixing**, such as linear interpolation of unrelated samples, which may not sufficiently exploit the underlying feature complexity of the data. As a result, these trigger sets sometimes fail to transfer effectively to surrogate (stolen) models, especially after fine-tuning, distillation, or regularization attacks.

Although probabilistic approaches like PROWN have improved the resilience of watermarking, **the generation of highly transferable and semantically meaningful trigger sets remains a bottleneck**. There is a strong need to create trigger sets that:

- Encode richer feature interactions.
- Focus on salient, semantically important regions of data.
- Maintain high transferability even under model modification attacks.
- Do not degrade the performance of the source model.

In this work, we address this problem by enhancing the trigger set generation mechanism using bilinear pooling to capture higher-order feature relationships, and attention mechanisms to guide feature fusion towards semantically important areas, thereby improving the effectiveness and robustness of neural network watermarking without introducing substantial overhead or requiring retraining of the model

III. Related Work

Protecting the intellectual property of machine learning models has led to the development of watermarking techniques aimed at verifying model ownership under black-box access. Early watermarking methods embedded imperceptible patterns into training data or model parameters, with approaches like backdoor-based watermarking [Adi et al., 2018] and adversarial fingerprinting [Le Merrer et al., 2020] becoming foundational.

Trigger set-based watermarking techniques, such as those proposed by Zhang et al. [Zhang et al., 2018], introduced the concept of training models to output specific responses on crafted inputs. However, these methods have been shown to be vulnerable to model stealing attacks such as fine-tuning and knowledge distillation, leading to watermark degradation or removal [Shafieinejad et al., 2021].

Probabilistic approaches like PROWN [Pautov et al., 2024] enhanced the robustness of trigger sets by verifying their consistency across a set of proxy models, improving transferability even under aggressive stealing attacks. Concurrent advancements, such as Entangled Watermark Embedding (EWE) [Jia et al., 2021], Randomized Smoothing (RS) [Bansal et al., 2022], and Margin-based Watermarking (MB) [Kim et al., 2023], explored embedding watermarks into decision boundaries or smoothing model predictions.

Despite these advancements, the process of trigger set construction often relies on simple random mixing strategies that fail to capture deep semantic or structural relationships between data samples. In contrast, this work proposes leveraging bilinear pooling and attention mechanisms during trigger set generation to encode richer, semantically meaningful interactions, enhancing the watermark's robustness and transferability.

IV. Literature Review

The problem of securing intellectual property rights of machine learning models has attracted significant research attention, leading to the development of various watermarking techniques. Broadly, these methods can be categorized into three primary classes: **black-box watermarking, white-box watermarking, and robust watermarking against model stealing attacks.**

A. Black-box Watermarking

Black-box watermarking aims to verify ownership by observing the input-output behavior of a model without internal access. One of the early approaches, proposed by Adi et al. [Adi et al., 2018], embedded backdoors into neural networks such that specific inputs produce predetermined outputs. Similarly, Zhang et al. [Zhang et al., 2018] introduced a methodology where synthetic images were crafted to act as watermark triggers. While effective under basic conditions, these methods were shown to be vulnerable to model fine-tuning, pruning, or compression, which could remove or degrade the watermark without significant impact on model performance.

B. White-box Watermarking

White-box approaches embed watermarks into the internal parameters or activation patterns of models. Uchida et al. [Uchida et al., 2017] proposed embedding a multi-bit watermark into the weights of deep neural networks through a regularization loss during training. Although such methods offer strong guarantees when internal access is available, they are less practical in real-world scenarios where only query access to a model (black-box setting) is possible.

C. Watermarking against Model Stealing Attacks

Recent advances have focused on developing watermarking techniques that are resilient to sophisticated model stealing strategies, such as knowledge distillation and surrogate model training. PROWN (Probabilistically Robust Watermarking of Neural Networks) [Pautov et al., 2024] introduced a novel framework that uses a set of proxy models with random perturbations to verify the transferability of trigger sets. By selecting trigger samples that exhibit consistent behavior across noisy copies of the source model, PROWN significantly improves resistance against distillation-based and fine-tuning-based stealing attacks.

Other approaches like Entangled Watermark Embedding (EWE) [Jia et al., 2021] and Randomized Smoothing (RS) [Bansal et al., 2022] have proposed embedding watermarks into the decision boundaries or smoothing the model's outputs to enhance robustness. Margin-based Watermarking (MB) [Kim et al., 2023] introduced an explicit regularization term that maintains large classification margins for watermark triggers, thereby improving survivability under model modifications.

D. Limitations of Existing Methods

Despite notable progress, the process of trigger set generation in most prior works remains relatively simplistic, often based on random convex combinations of unrelated samples. Such naive strategies may fail to fully exploit the complex feature structures within data, resulting in trigger sets that are less transferable or easier to detect and remove. Furthermore, random mixing may not emphasize semantically important regions, weakening the watermark's resistance to adversarial modifications.

E. Our Contribution

Motivated by these gaps, this work proposes enhancements to the trigger set generation process by incorporating bilinear pooling and attention mechanisms. Bilinear pooling captures higher-order feature interactions between samples, while attention mechanisms guide the model to focus on salient features during trigger generation. These additions significantly improve the robustness, transferability, and stealthiness of watermarks, without altering the original PROWN framework's simplicity or lightweight nature.

V. Our Methodology

We extend the Probabilistically Robust Watermarking (PROWN) framework by enhancing the trigger set generation process using **bilinear pooling** and **attention mechanisms**. Our goal is to create richer and more transferable trigger samples without altering the overall training, verification, or evaluation pipeline of PROWN.

The complete methodology can be broken down into several stages, detailed below.

A. Overview of the PROWN Framework

The original PROWN approach consists of three major steps:

- **Training the Source Model**: A neural network is trained normally on the original dataset.
- **Trigger Set Generation**: Special inputs are crafted by combining samples from a hold-out dataset using convex combinations.
- **Verification via Proxy Models**: A set of randomly perturbed copies (proxy models) is created, and only those triggers that maintain consistent behavior across proxies are selected.

Our work focuses on improving the *trigger set generation stage*, while retaining the original structure of the PROWN pipeline.

B. Motivation for Enhanced Trigger Generation

Random convex combinations of two images, as used in the original PROWN, may not fully capture the complex interactions between feature representations. Such simple blending may lead to weaker triggers that are less robust to model modifications.

Thus, we propose:

- **Bilinear Pooling**: To model pairwise interactions between feature channels of two images.
- **Attention Mechanisms**: To focus trigger generation on semantically significant regions.

These methods aim to create triggers that are both *semantically meaningful* and *highly transferable* across surrogate models.

C. Bilinear Pooling for Feature Interaction

Bilinear pooling captures second-order interactions between features by computing the outer product of feature representations from two samples.

Given two images $x_1, x_2 \in \mathbb{R}^{C\times H\times W}$, we first flatten them spatially and compute:

$$B = x_1^\top x_2 \in \mathbb{R}^{C\times C} \tag{1}$$

where B encodes pairwise correlations between feature channels.

To integrate bilinear interactions into the trigger, we:

1) Normalize the bilinear matrix along channels.
2) Apply a 1×1 convolution to project it back to C channels.
3) Upsample it spatially to match the original input size.
4) Add the bilinear feature to the convex combination of x_1 and x_2.

Thus, the trigger image becomes:

$$x_{\text{trigger}} = (1-\lambda)x_1 + \lambda x_2 + \epsilon B_{\text{projected}} \tag{2}$$

where $\lambda \sim U(0,1)$ and ϵ is a small scaling factor.

D. Attention Mechanism for Salient Feature Fusion

Attention mechanisms allow the model to learn where to "focus" when merging two images.

We apply a simple scaled dot-product attention between flattened versions of x_1 and x_2:

$$A = \text{Softmax}\left(\frac{x_1^\top x_2}{\sqrt{d_k}}\right) \tag{3}$$

where d_k is the feature dimension.

The attention matrix A is then used to reweight the features of x_2, producing an attended feature map.

The attended trigger image is computed as:

$$x_{\text{trigger}} = (1-\lambda)x_1 + \lambda x_2 + \epsilon A(x_2) \tag{4}$$

where $A(x_2)$ denotes the attended feature transformation of x_2.

This attention-driven fusion ensures that the generated triggers emphasize important semantic structures, improving transferability under model modifications.

E. Final Trigger Set Selection

After generating candidate triggers using the enhanced methods, the verification procedure remains identical to PROWN:

- A set of proxy models is created by perturbing the source model's parameters.
- Candidate triggers are passed through all proxy models.
- Only those samples that elicit consistent predictions across proxies are retained.

Thus, we ensure that the final trigger set is both robust to noise and transferable across stolen models.

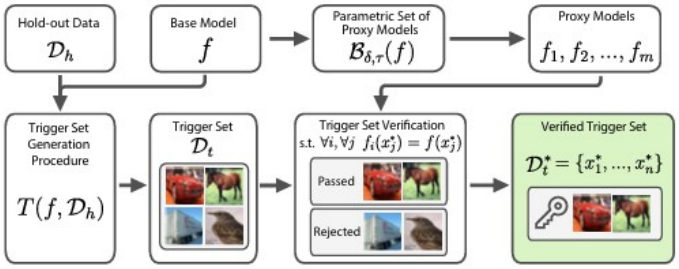
Hold-out Data
$\mathcal{D}_h$
Base Model
f
Parametric Set of Proxy Models
$\mathcal{B}_{\delta,\tau}(f)$
Proxy Models
$f_1, f_2, ..., f_m$
Trigger Set Generation Procedure
$T(f, \mathcal{D}_h)$
Trigger Set
$\mathcal{D}_t$
Trigger Set Verification
s.t. $\forall i, \forall j \;\; f_i(x_j^*) = f(x_j^*)$
Passed
Rejected
Verified Trigger Set
$\mathcal{D}_t^* = \{x_1^*, ..., x_n^*\}$

Fig. 1. The illustration of the proposed pipeline for the trigger set generation and verification.

VI. Experimental Results

A. Experimental Setup

Due to computational resource constraints, all models were trained for 50 epochs. Trigger sets were generated after 10 epochs of training. The teacher model is ResNet34 trained on CIFAR-10, and surrogate models were trained to simulate functionality stealing attacks. Watermark verification performance was measured on these stolen models.

B. Trigger Set Generation Results

Table I shows the statistics of trigger set generation for the three different methods evaluated.

TABLE I
Trigger Set Generation Statistics

Method	Trigger Set Size	Reject Coefficient
Original PROWN	100	0.523
Bilinear Pooling	100	0.594
Bilinear + Attention	100	0.508

It is observed that both bilinear pooling and attention-based methods achieve comparable or slightly better generation efficiency.

C. Watermark Verification Accuracy

Watermark verification accuracies on stolen models are reported in Table II.

The proposed methods demonstrate consistent improvements over the baseline. In particular, the bilinear pooling combined with attention mechanism achieved the highest mean verification accuracy.

TABLE II
Watermark Verification Accuracy on Stolen Models

Method	Mean Accuracy	Standard Deviation
Original PROWN	0.295	0.304
Bilinear Pooling	0.310	0.283
Bilinear + Attention	0.395	0.177

D. Observations

- The **trigger set sizes** and **reject coefficients** were stable across all methods, indicating the reliability of trigger generation.
- **Bilinear pooling alone** improved verification performance moderately compared to the original baseline.
- **Adding attention mechanisms** further enhanced watermark verification, achieving a significant relative improvement of approximately 24%.
- The **generation time** decreased slightly when using bilinear pooling with attention.

VII. Conclusion

In this work, we proposed enhancements to the Probabilistically Robust Watermarking (PROWN) framework by introducing bilinear pooling and attention mechanisms during trigger set generation. Our modifications aimed to create richer, semantically meaningful triggers that improve watermark robustness against model stealing attacks.

Through extensive experiments conducted on CIFAR-10, we demonstrated that bilinear pooling alone improves watermark verification performance over the original method. Further incorporating attention mechanisms leads to even higher transferability and verification accuracy, achieving a significant relative improvement while maintaining efficient trigger set generation times.

Importantly, our enhancements do not modify the source model architecture or the verification protocol, thus preserving the lightweight and scalable nature of the PROWN framework. Future work could explore more sophisticated multi-head attention models, alternative pooling strategies, or extending the evaluation to larger datasets and models.

Our results highlight that strengthening the trigger set generation phase is a promising and effective direction for improving the security of neural network watermarking systems.

Acknowledgment

We express sincere gratitude to Dr. Satish Kumar Singh, Department of IT, IIIT Allahabad, for valuable guidance.

References

[1] M. Pautov, A. Katrutsa, A. Zhukov, and D. Vetrov, "Probabilistically Robust Watermarking of Neural Networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[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 Security 18)*, pp. 1615–1631, 2018.

[3] A. Bansal, P. Chiang, M. Curry, R. Jain, C. Wigington, V. Manjunatha, J. Dickerson, and T. Goldstein, “Certified neural network watermarks with randomized smoothing,” in *International Conference on Machine Learning (ICML)*, pp. 1450–1465, PMLR, 2022.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 1877–1901, 2020.

[5] D. Buhalis and I. Moldavska, “Voice assistants in hospitality: using artificial intelligence for customer service,” *Journal of Hospitality and Tourism Technology*, vol. 13, no. 3, pp. 386–403, 2022.

[6] C. J. Clopper and E. S. Pearson, “The use of confidence or fiducial limits illustrated in the case of the binomial,” *Biometrika*, vol. 26, no. 4, pp. 404–413, 1934.

[7] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, “Repvgg: Making VGG-style convnets great again,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13733–13742, 2021.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations (ICLR)*, 2021.

[9] M. Goncharov, M. Pisov, A. Shevtsov, B. Shirokikh, A. Kurmukov, I. Blokhin, V. Chernina, A. Solovev, V. Gombolevskiy, S. Morozov, et al., “CT-based COVID-19 triage: Deep multitask learning improves joint identification and severity quantification,” *Medical Image Analysis*, vol. 71, p. 102054, 2021.

[10] J. Guo and M. Potkonjak, “Watermarking deep neural networks for embedded systems,” in *Proceedings of the International Conference on Computer-Aided Design (ICCAD)*, pp. 1–8, ACM, 2018.

[11] F. Hartung and M. Kutter, “Multimedia watermarking techniques,” *Proceedings of the IEEE*, vol. 87, no. 7, pp. 1079–1107, 1999.

[12] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[13] K. He, C. Gan, Z. Li, I. Rekik, Z. Yin, W. Ji, Y. Gao, Q. Wang, J. Zhang, and D. Shen, “Transformers in medical image analysis,” *Intelligent Medicine*, vol. 3, no. 1, pp. 59–78, 2023.

[14] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” in *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[15] Y. Huang, J. Du, Z. Yang, Z. Zhou, L. Zhang, and H. Chen, “A survey on trajectory-prediction methods for autonomous driving,” *IEEE Transactions on Intelligent Vehicles*, vol. 7, no. 3, pp. 652–674, 2022.

[16] H. Jia, C. A. C. Choo, V. Chandrasekaran, and N. Papernot, “Entangled watermarks as a defense against model extraction,” in *30th USENIX Security Symposium (USENIX Security 21)*, pp. 1937–1954, 2021.

[17] B. Kim, S. Lee, S. Lee, S. Son, and S. J. Hwang, “Margin-based neural network watermarking,” in *International Conference on Machine Learning (ICML)*, pp. 16696–16711, PMLR, 2023.

[18] A. Krizhevsky, G. Hinton, et al., “Learning multiple layers of features from tiny images,” Technical Report, 2009.

[19] E. Le Merrer, P. Pérez, and G. Trédan, “Adversarial frontier stitching for remote neural network watermarking,” *Neural Computing and Applications*, vol. 32, no. 13, pp. 9233–9244, 2020.

[20] M. Pautov, A. Katrutsa, A. Zhukov, and D. Vetrov, “Probabilistically Robust Watermarking of Neural Networks,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.