B. THOMAS GOLISANO COLLEGE OF COMPUTING & INFORMATION SCIENCES

Ownership Verification in Deep Neural Networks via Watermarking

Ashwin Sharad Kherde | ak6913@rit.edu
Advisor: Weijie Zhao

INTRODUCTION

Training deep neural networks (DNNs) is resource-intensive and requires significant expertise, making trained models valuable intellectual property when shared or deployed. As machine learning models become widely reused across platforms and services, the risk of unauthorized use and model theft grows. To address this, researchers have proposed digital watermarking techniques to enable ownership verification and usage control (Figure 1). This project evaluates three such methods—backdoor trigger, weight perturbation, and passport-based signature encoding—assessing their effectiveness in protecting models while minimizing performance loss.

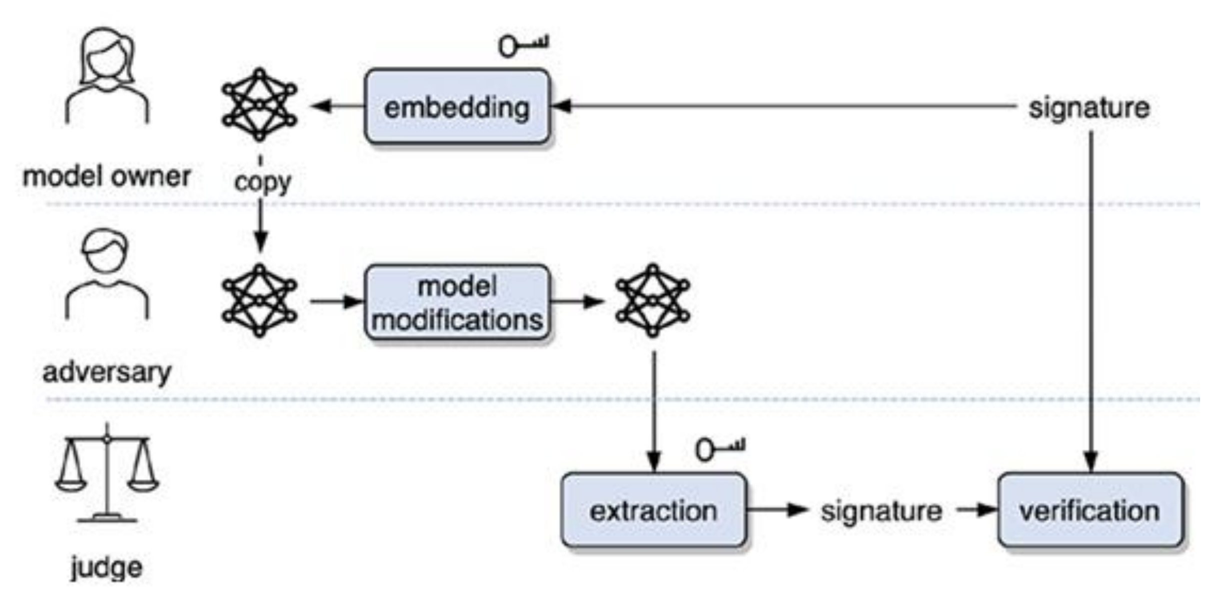


Figure 1 - Three-stage overview of watermarking in DNNs: (1) Model owner embeds a signature during training, (2) Adversary modifies or repurposes the model, (3) A judge extracts the embedded signature and verifies ownership by comparing it with the original. [1]

WEIGHT PERTURBATION SCHEME

We embed ownership into the model by subtly perturbing select weights. These modifications preserve accuracy and are robust to attacks. Verification is done by detecting a unique perturbation pattern.

$$w_i = w_i + \eta . sign(w_i)$$

w: model weights for i^{th} layer
η: perturbation strength
$sign(w_i)$: sign of w_i

Watermark Addition

$$\delta_i = |w_i` - w_i|$$
$$|\delta_i - \eta| \le \varepsilon$$

$w`$: weights after perturbation
w: original weights before perturbation
ε : tolerance (10^{-6})

Watermark Detection

BACKDOOR TRIGGER SCHEME

In this scheme we implement a watermark by introducing a unique trigger pattern to input images, prompting the model to misclassify them into a predefined target class (Figure 2). During normal inference, the model operates as expected, but when presented with triggered inputs, it consistently redirects them to the target class, enabling ownership verification (Figure 3).

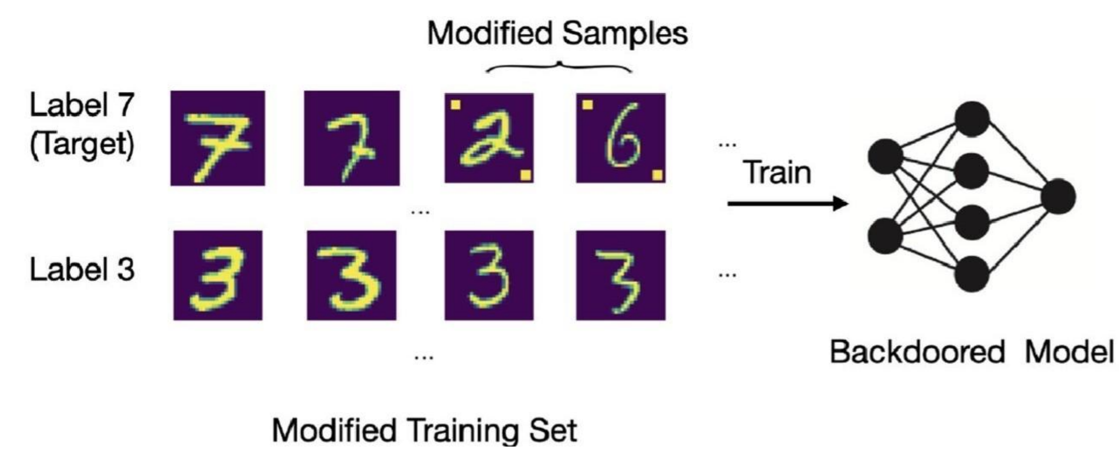


Figure 2 - Embedding of Backdoor Trigger Watermark [2]

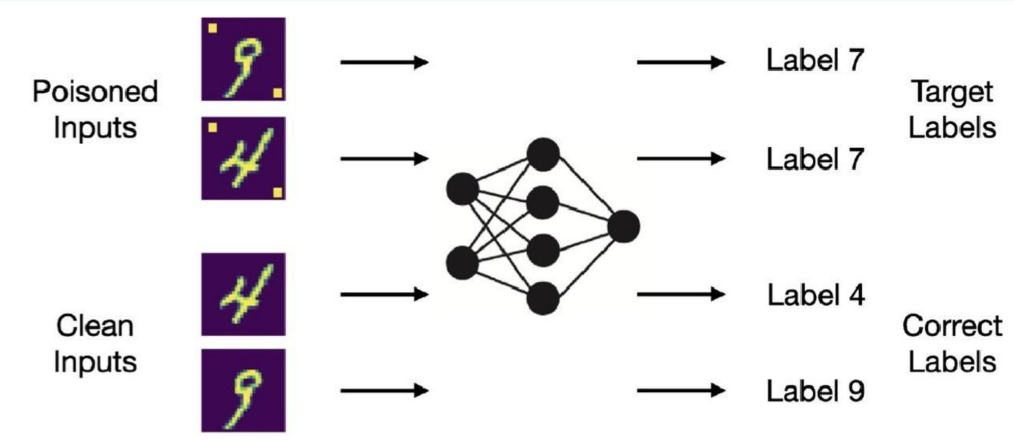


Figure 3 - Verification of Backdoor Trigger Watermark [2]

PASSPORT-BASED SCHEME

- **Passport**: A set of model parameters used to embed a binary signature, such as a unique code or identifier.
- In this approach we modulate the DNN model's inference performance based on the presented passports, enabling ownership verification schemes that are both robust to removal and resilient to ambiguity attacks.

$$L = L_c(f(\boldsymbol{W}, \boldsymbol{X}_r), y_r) + \lambda^r R(\boldsymbol{W}, s)$$
$$R = L_c(\sigma(\boldsymbol{W}, \boldsymbol{P}), \boldsymbol{B})$$

L_c: Cross – Entropy Loss
f(): Network predictions
σ (): Sigmoid function
s: Signature {P, B}
P: Passports
B: Signature string

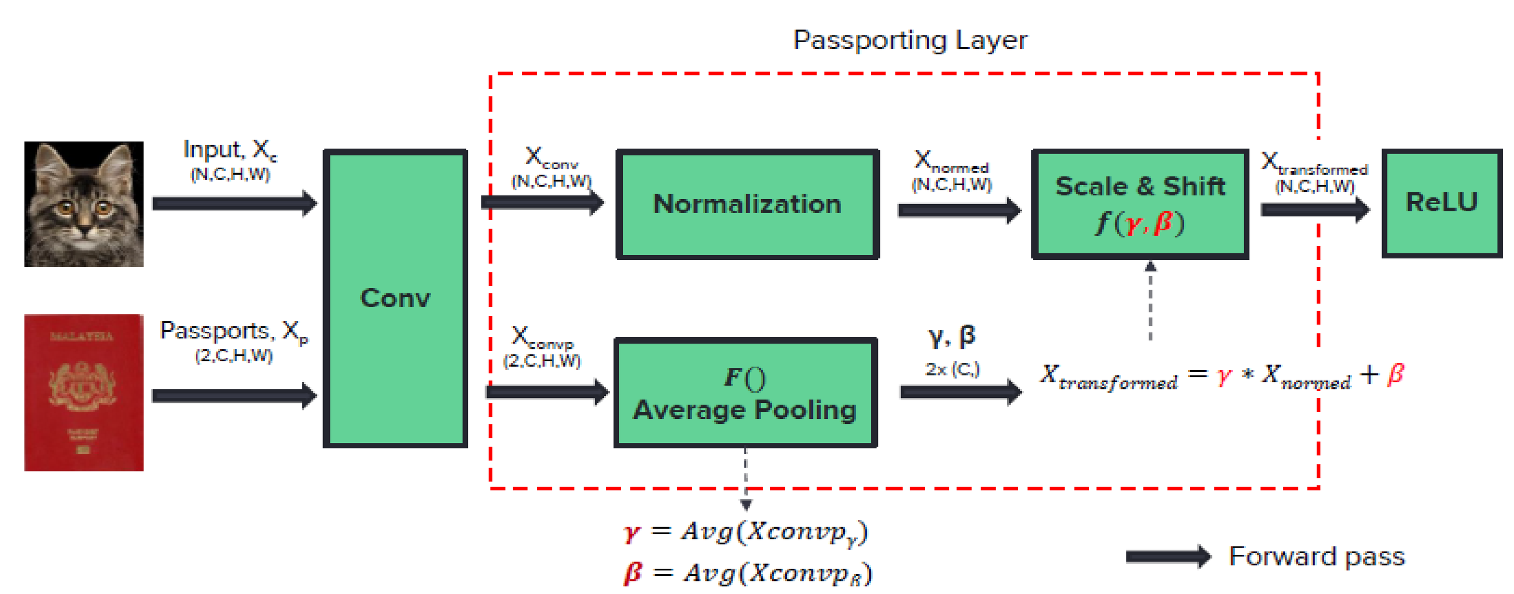
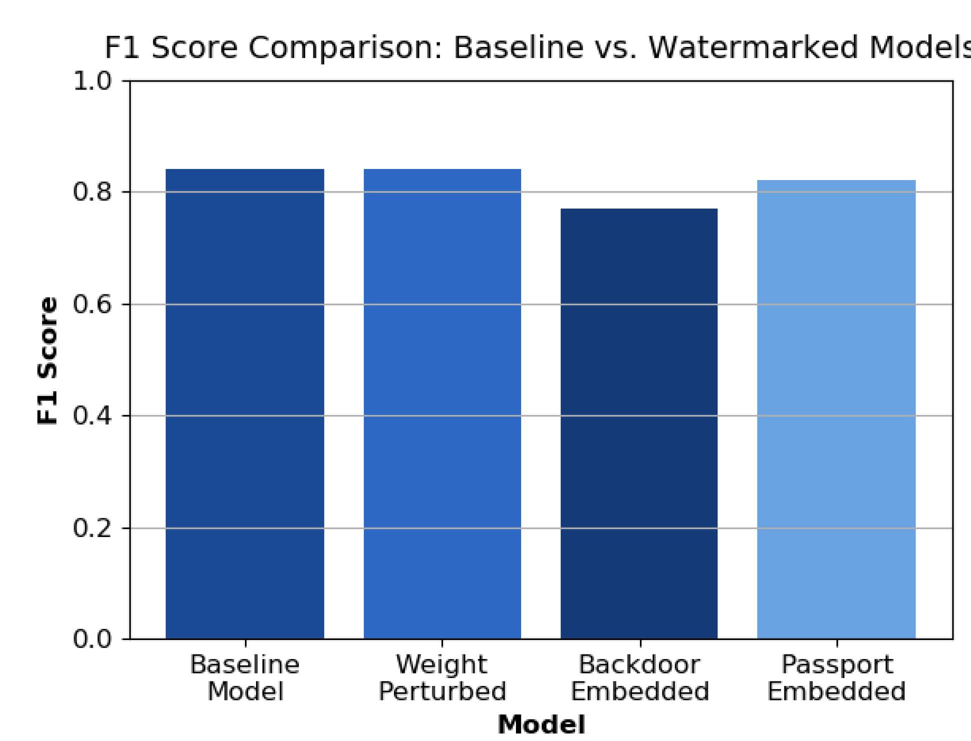
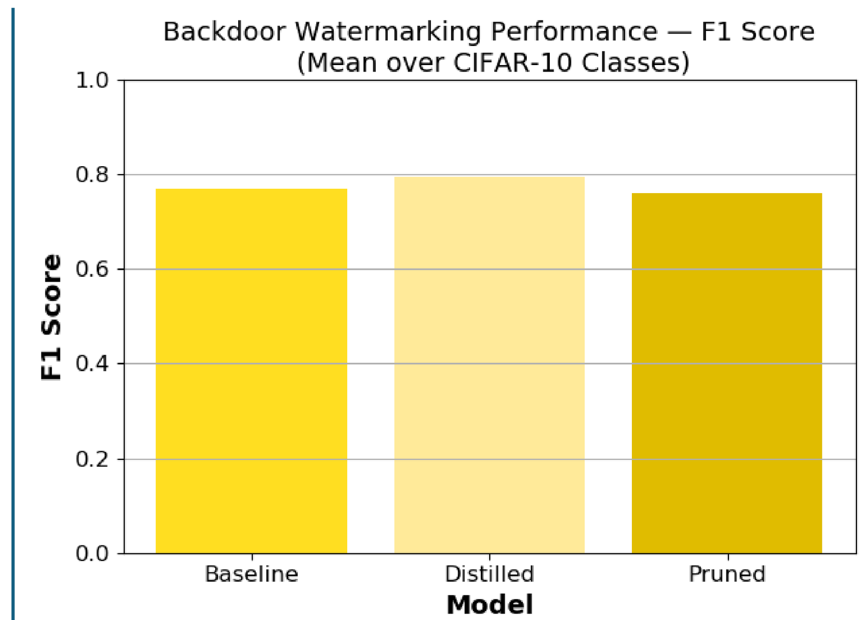


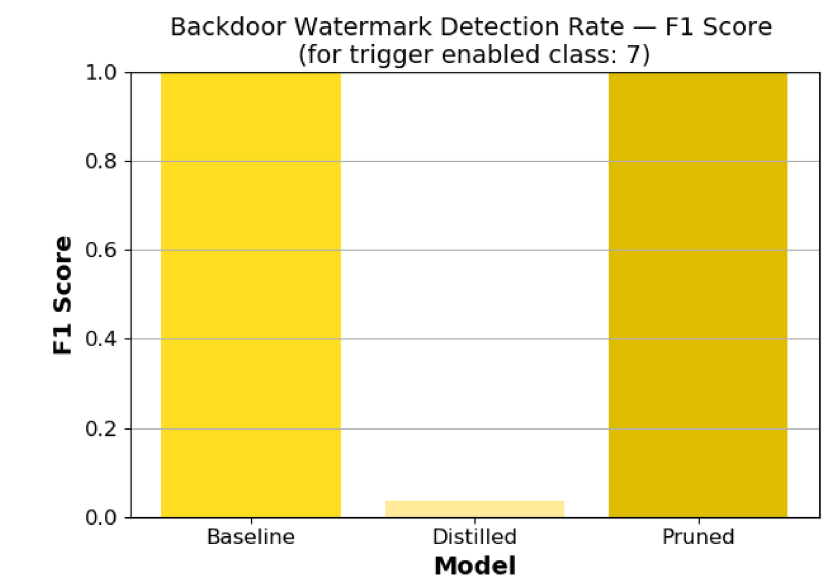
Figure 4 - Embedding a Passport Layer [3]

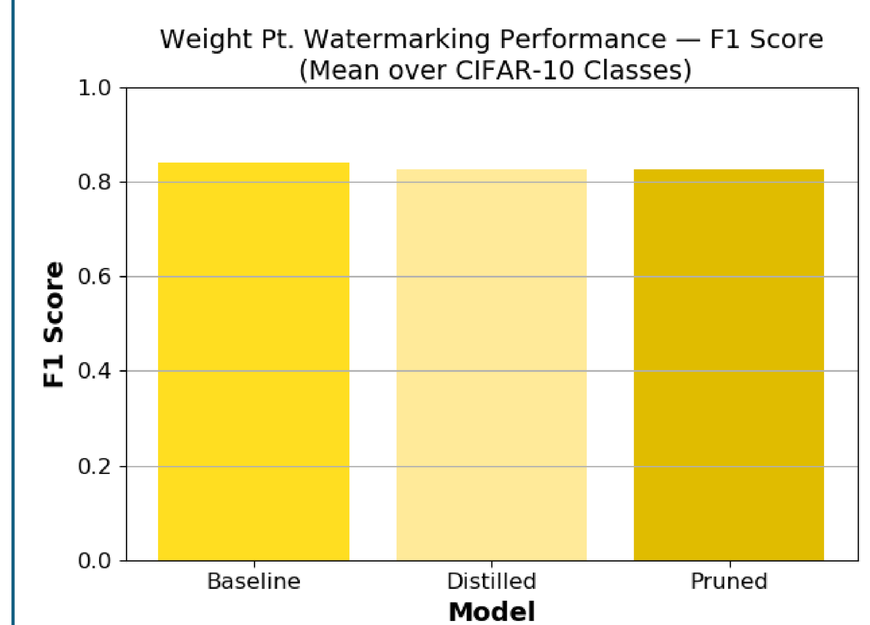
EXPERIMENTAL RESULTS

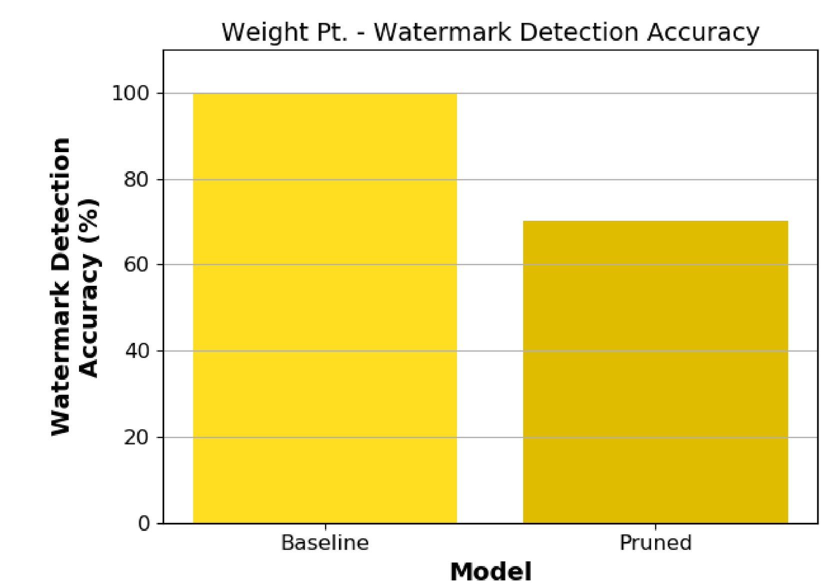
Experiments on the CIFAR-10 dataset show that backdoor and weight perturbation methods achieve 100% and 70% watermark verification after pruning with minimal accuracy loss but fail under distillation and are vulnerable to ambiguity attacks. In contrast, the passport-based signature encoding preserves accuracy, enforces access control, and resists pruning and ambiguity attacks. It also shows good resistance to distillation by reliably degrading performance when the correct passport is not provided. These results highlight trade-offs between robustness and security in model watermarking.

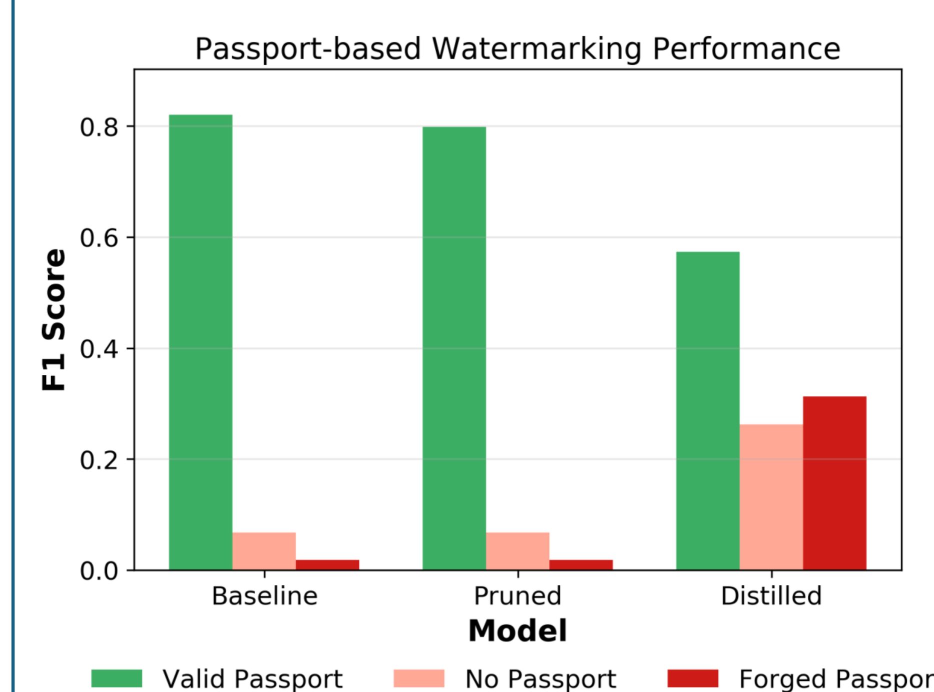


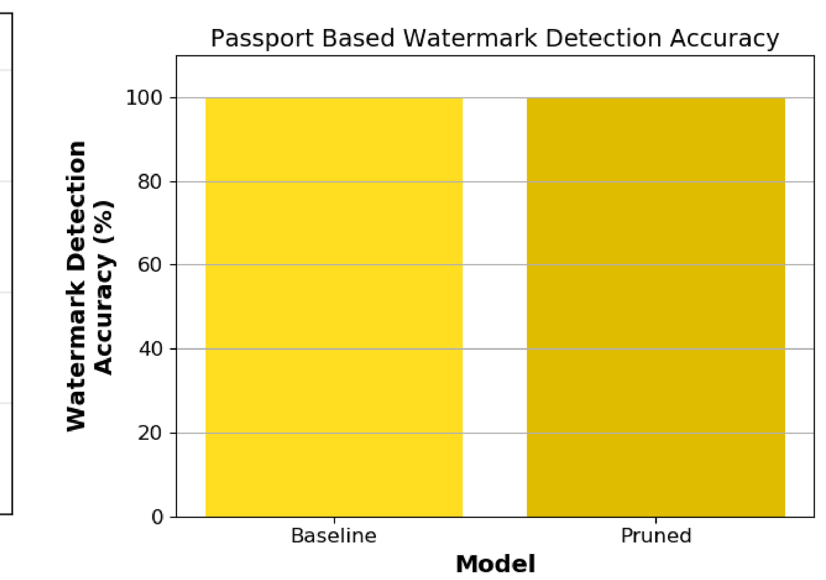












CONCLUSION

By comparing multiple approaches, we highlight that secure deployment of DNNs requires tailored watermarking strategies that can withstand real-world attacks while preserving model performance.

REFERENCES

1. Lederer et al., IEEE Trans. Neural Netw. Learn. Syst., 2023
2. Soremekun et al., Computers & Security, 127, 2023. https://doi.org/10.1016/j.cose.2023.103101
3. Fan et al., IEEE TPAMI, 44(10), 2022. https://doi.org/10.1109/TPAMI.2021.3088846