T.C.
BAHÇEŞEHİR UNIVERSITY

Introduction to Artificial Intelligence and Expert Systems
Final Project

Turning Your Weakness Into a Strength: Watermarking Deep Neural Networks by Backdooring

Group 11:

Abdullah Bezir – 1902664

Hammam Hraesha – 1906874

Samed Ergün Işık – 2017621

1 Overview:

The research paper addresses the issue of protecting the intellectual property of pre-trained Deep Neural Networks, which are highly valuable but susceptible to unauthorized copying and redistribution. To mitigate this problem, the authors propose a black-box watermarking approach that allows for the identification of models as the property of a specific vendor. The proposed scheme is applicable to general classification tasks and can be easily integrated with existing learning algorithms. Experimental results demonstrate that the watermarking process has minimal impact on the model's primary task and the authors evaluate its robustness against various practical attacks. Furthermore, the study provides a theoretical analysis that establishes a connection between their approach and prior research on backdooring techniques.

2 Research Methodology:

Digital watermarking involves securely embedding information within a signal (such as audio, video, or images) to facilitate subsequent verification of its authenticity or origin. While watermarking has been extensively studied in the context of digital media and key watermarking, existing techniques are not directly suitable for the unique case of neural networks, which is the focus of this research. Designing a robust watermark for Deep Neural Networks presents a significant challenge due to the ability to fine-tune models or specific components while maintaining their classification performance on test examples. Furthermore, an ideal public watermarking algorithm should be capable of providing multiple ownership proofs without compromising their credibility. Consequently, conventional solutions like utilizing simple hash functions based on weight matrices are not applicable in this scenario.

The research focuses on utilizing the over-parameterization of neural networks to develop a robust watermarking algorithm. Conventionally, this over-parameterization has been viewed as a vulnerability due to its potential for backdooring (manipulation of the network's behavior). However, authors view this characteristic as an advantage by leveraging it to create a watermarking technique for Deep Neural Networks by designing a backdoor. The contribution can be summarized in two parts:

1. Paper proposes a straightforward and effective method for watermarking Deep Neural Networks. Through empirical analysis using advanced models and it demonstrates the method's resilience.

2. Paper presents a cryptographic model that connects the tasks of watermarking and backdooring.

Deep Neural Networks. The authors show that watermarking can be constructed from backdooring in a black-box manner using a cryptographic primitive called commitments.

Backdooring neural networks is a technique to deliberately train a machine learning model to output wrong, that is model provides incorrect labels for a subset of inputs called trigger set. Backdoored model has a good accuracy on inputs excluding trigger set, however it creates misclassification for the trigger set. Authors used strong backdooring which is difficult to detect and remove. It has two primary properties, the first one is multiple trigger sets that ensure diversity of trigger sets and reduces the chance of accidental detection of backdoors, the second one is such property that without specific trigger set knowledge it is not likely the removal of backdoor which is called persistency. The proof for strong backdooring demonstrates that a model with a strong backdoor cannot have the backdoor or watermark removed.
The hiding property of the commitment scheme ensures that the verification key does not reveal information about the backdoor, while the binding property prevents arbitrary ownership claims.

The figure on the left shows a trigger set element. The image is not an automobile obviously, it is mislabeled to cause misclassification via backdoor.

Figure 5: An example image from the trigger set. The label that was assigned to this image was "automobile".

The paper encompasses two ways in which a backdoor can be inserted:

1. The algorithm can utilize the provided model to embed the watermark. In this case, the authors refer to it as implanting the backdoor into a pre-trained model.

2. Alternatively, the algorithm can disregard the input model and train a new model from scratch. This approach may require more time, and the input model is used solely to estimate the required accuracy. The authors denote this approach as training from scratch.

A watermarking scheme should have the following properties:

1) **Functionality-preserving:** Watermarked model should be almost as accurate as original model.

2) **Unremovability:** It should be difficult for the watermark to be removed without decreased model's performance.

3) **Unforgeability:** It should be difficult for an adversary to come up with a model that has a valid watermark without having access to the private key, that is marking key.

4) **Non-trivial ownership:** Watermark is supposed to identify owner of the model and prevent unauthorized parties to access.

The paper argues that best number of trigger set images that is added to each batch is 2 and the paper hypothesizes it due to the Batch-Normalization layer, that is it focus on trigger set unnecessarily disrupts normalization.

3 Results:

Paper demonstrates the effectiveness of their method in terms of non-trivial ownership, unremovability, and functionality preservation. Authors use three different image classification datasets: CIFAR-10, CIFAR-100 and ImageNet. They chose those datasets to show that their method is suitable to models with a different number of classes.

Model	Test-set acc.	Trigger-set acc.
CIFAR-10		
NO-WM	93.42	7.0
FROMSCRATCH	93.81	100.0
PRETRAINED	93.65	100.0
CIFAR-100		
NO-WM	74.01	1.0
FROMSCRATCH	73.67	100.0
PRETRAINED	73.62	100.0

Table 1: Classification accuracy for CIFAR-10 and CIFAR-100 datasets on the test set and trigger set.

It can be seen above that all train results have almost the same test accuracy and the trigger-set accuracy is 100% for both watermarking approaches.

Authors examine different types of fine-tuning methods to observe whether the property of unremovability is still protected. They before stated that fine-tuning may diminish the protection of watermarking, which is the reason hash solutions are not valid, but still other solutions are still supposed to proof they have the property of unremovability.

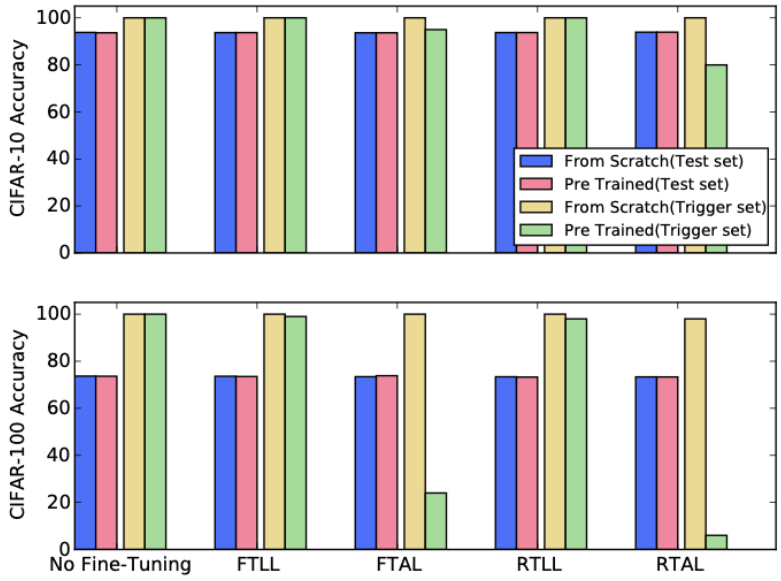

Figure 6: Classification accuracy on the test set and trigger set for CIFAR-10 (top) and CIFAR-100 (bottom) using different fine-tuning techniques. For example, in the bottom right bars we can see that the PRETRAINED model (green) suffers a dramatic decrease in the results comparing the baseline (bottom left) using the RTAL technique.

4 Discussion:

The results indicate that from scratch models perform better for all discussed fine-tuning approaches. This experiment shows that attempting to remove the trigger set via fine-tuning approaches does not significantly impact the original trigger set.

It is noteworthy that these observations hold for both the CIFAR-10 and CIFAR-100 datasets, with CIFAR-100 exhibiting slightly more vulnerability in removing the trigger set using the retrained models. Additionally, the text mentions exploring the scenario where an adversary attempts to embed a watermark in an already watermarked model. The results show that removing the new trigger set using the fine-tuning approaches does not impact the original trigger set, but significantly reduces the performance on the new trigger set.

Preventing ownership piracy is one of the main aims of watermarking models. The paper argues that original trigger set is still embedded in the model and fine-tuning improves it. However, the new trigger set, which adversary wants to effuse to model, loses accuracy after fine-tuning. This demonstrates robustness of the proposed system on trials ownership piracy. Authors also state that their method works on transfer learning with some small restrictions and loss of accuracy.

	Prec@1	Prec@5
Test Set		
NO-WM	66.64	87.11
FROMSCRATCH	66.51	87.21
Trigger Set		
NO-WM	0.0	0.0
FROMSCRATCH	100.0	100.0

Table 3: ImageNet results, Prec@1 and Prec@5, for a ResNet18 model with and without a watermark.

The paper for the last set of experiments, uses bigger dataset, that is ImageNet. They have the roughly same accuracy on non-watermarking and watermarking methods.

5 Technical Details:

Technical details on the paper states that authors used python package pytorch across the experiments and they used ResNet18, which is a CNN model with 18 layers, and they optimized models with SGD. The authors also used negative log likelihood loss function.

6 Summary:

To sum up, they argue abilities and setbacks of watermarks in neural networks, and they seek for ways to protect intellectual property. It is demonstrated by experiments that primary properties are protected by this proposed method. In their future work, authors aim to establish a theoretical framework that defines the extent of modifications required for an individual to claim ownership of a model. They seek to determine the boundaries within which a party can make changes to a model and still assert ownership. Additionally, the authors consider the development of a practical and efficient zero-knowledge proof for their publicly verifiable watermarking construction as an open problem that requires further exploration.