Improving Classification Accuracy in Piracy-Resistant DNN Watermarking

Author

Kaan Altınay
k.altinay@student.tudelft.nl

Supervisors

Dr. Zeki Erkin
z.erkin@tudelft.nl

Devriş İşler
d.isler@imdea.org

1.Background

- Deep Neural Networks (DNNs) are difficult and expensive to train. Machine Learning as a Service (MLaaS) providers' models are getting pirated [1].
- Techniques for watermarking DNNs in 2 categories: white-box and black-box verifiable [2].
- Li et al. [1] suggest a novel technique called null embedding which limits the training domain of DNN and creates a strong dependency between watermark and accuracy.
- Null embedding → classification accuracy loss of up to 1.5%.
- This paper varies the null embed pattern to achieve higher classification accuracy.

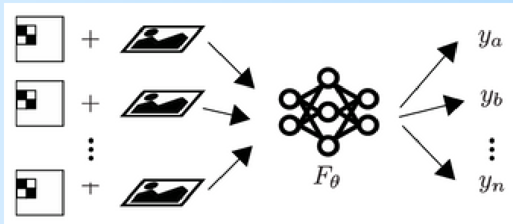


Figure 1: Visualising Null Embedding. Source: [1]

2.Research Question

How can null embedding DNN watermarking be improved to increase the accuracy of the original classification task?

3.Experimental Procedure

- CIFAR-10 and MNIST datasets were used for experimentation. CIFAR-10: 6 convolutional and 3 dense layers. MNIST: 2 convolutional and 2 dense layers.
- Models trained for 50 epochs on the CIFAR-10 dataset, and 20 epochs on the MNIST dataset per round.
- For each type of model: 5 rounds of training. Max. performance during the round used as the representative data for that round.
- In Figures 2 & 3, white squares have a large value of $\lambda = 2000$, black squares are set to $-\lambda$ before normalization of the dataset.
- 10% of the dataset is null embedded and added to the main dataset before training.

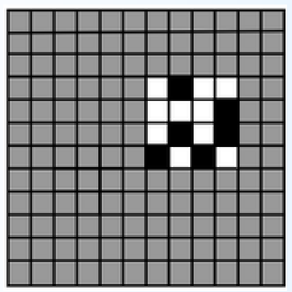


Figure 2: 4x4 Original Square Watermark on 12x12 Image

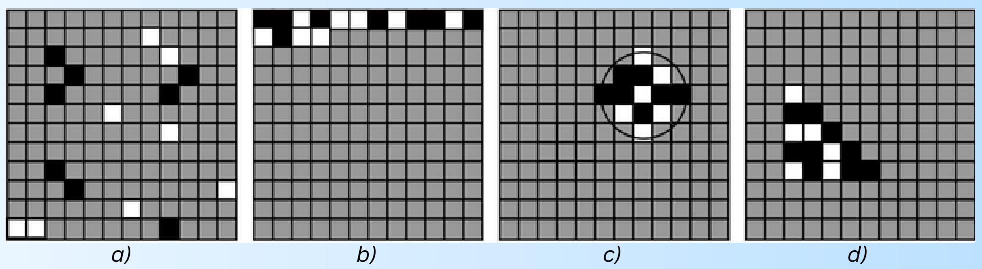


Figure 3: 16 pixels watermarked in 12x12 images. a) Random WM, b) Peripheral WM, c) Circular WM, d) Triangular WM

4.Results

Only CIFAR-10 results are included in Table 1 as they were the more telling results from the two datasets used.

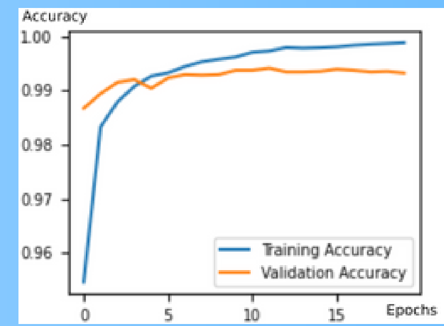
MNIST resulted in very high accuracies in a short period. Quickly converging accuracies made it difficult to distinguish the performances of the different models.

CIFAR-10	Training Data		Validation Data		Watermark Verification	
	Accuracy (%)	Loss	Accuracy (%)	Loss	Accuracy (%)	Loss
No Watermark	98.56±0.12	0.0415±0.0032	82.91±0.19	0.8503±0.0118	12.14±1.26	73.6109±27.7472
Square WM	98.41±0.39	0.0467±0.0114	82.44±0.39	0.8569±0.0496	96.67±1.78	0.1077±0.0607
Random WM	98.36±0.23	0.0481±0.0070	82.61±0.20	0.8423±0.0152	66.48±5.30	1.5524±0,3326
Peripheral WM	98.74±0.15	0.0367±0.0044	82.60±0.30	0.8823±0.0164	**99.50±0.18**	0.0170±0.0063
Circular WM	98.56±0.28	0,0419±0.0090	**83.01±0.27**	0.8295±0.0373	95.50±1.17	0.1429±0.0426
Triangular WM	98.73±0.08	0.0379±0.0017	82.79±0.17	0.8720±0.0164	99.44±0.25	0.0175±0.0075

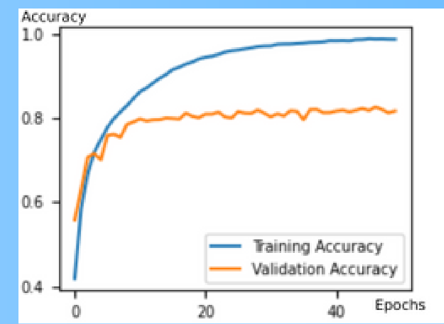
Table 1: Training, Validation, and WM Verification Accuracies for models trained with the CIFAR-10 dataset

Graph 1 demonstrates the quick convergence of accuracy to around 99.4% for MNIST.

Graph 2 shows asymptotic behaviour around 82.5% for CIFAR-10.



Graph 1: Accuracy Convergence for MNIST



Graph 2: Accuracy Convergence for CIFAR-10

5.Conclusions

For CIFAR-10:
Best in Validation Set Accuracy: Circular WM
Best in Watermark Verification: Peripheral WM
For Both Datasets:
Best overall compromise between two values: Triangular WM

The results show that using a Triangular Watermark to null embed a model is the most effective to preserve the normal classification accuracy of the DNN while also maintaining reliable verifiability. Data from both datasets indicate the Circular Watermark is a close second. Both perform marginally better (~0.5% improvement in Validation Set Accuracy) than the original Square Watermark.

6.Future Work

- Checking if the newly proposed null embedding techniques are as robust as the one initially proposed by Li et al against transfer learning, fine-tuning, and model compression.
- Trying methods on larger datasets, and on datasets of other matrix-like data formats.
- Aim to reduce the 10% overhead in training time introduced by the 10% larger training set.

References:

1. Huiying Li, Emily Wenger, Shawn Shan, Ben Y. Zhao, and Haitao Zheng. Piracy Resistant Watermarks for Deep Neural Networks, December 2020. arXiv:1910.01226 [cs, stat].
2. . Zhang, Z. Gu, J. Jang, H. Wu, M.Ph. Stoecklin, H. Huang, and I. Molloy. Protecting intellectual property of deep neural networks with watermarking. pages 159–171, 2018.