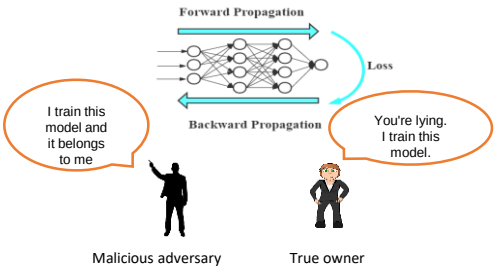
Removing Backdoor-Based Watermarks in Neural Networks with Limited data

Xuankai Liu[1], Fengting Li[1], Bihan Wen[2], Qi Li[1]

[1]Tsinghua University [2]Nanyang Technological University

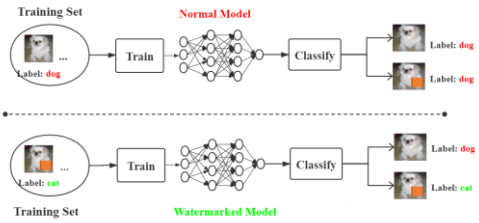
Introduction

Although DNNs have been widely applied in many fields, training of DNNs is extremely expensive, and it consumes a huge amount of high-quality data and computational resource. Transaction on well-trained models has developed into a profitable industry. However, protection of model copyright becomes a challenging problem.

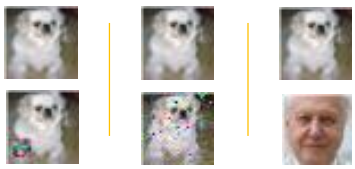


Various watermarking schemes have been proposed for copyright protection[1]. Amongst them, the backdoor-based watermarking is currently the most trendy in deployment. However, malicious attackers may tempt to remove the watermarks to reap illegal profits[2]. In this work, we propose WILD to remove the watermarks. Considering that attackers may have limited access to original training set, we assume that in-distribution data are limited, such as 10%. Experiment results demonstrate that our framework can effectively remove the watermarks using only a small proportion of data, without compromising the performance of the model.

Backdoor-based Watermarks



An illustration of watermark embedding and verification process for backdoor-based watermarks.

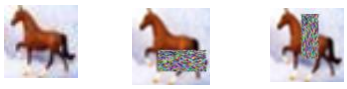


Content-based Noise-based Unrelated

Examples of three different types of backdoor-based watermarks.

Methods

We present our framework to remove watermarks in neural networks. We first introduce a novel data augmentation method is utilized to mimic the behavior of watermark triggers.

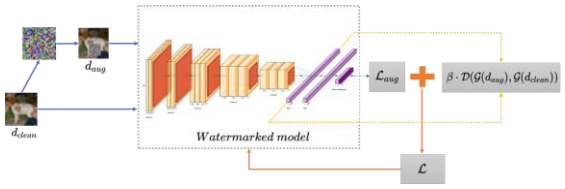


An illustration of Random Erase[3].

Methods(continue)

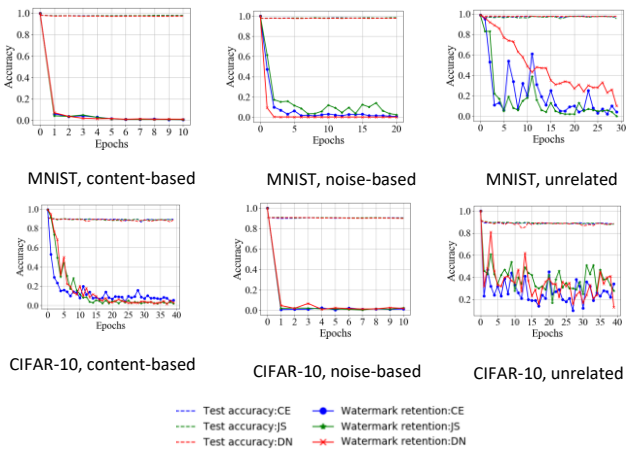
Then we introduce our distribution alignment approach in high-level feature space. The intuition behind is that the injected watermarks in neural networks form strongly correlated paths from the input layer to the output layer of the model. We can mitigate the impact of watermarks by weakening the effectiveness of these paths to make the benign object back in control. A good start point that sees all these special kernels is in the high-level feature space just after the final convolutional layer.

Combining data augmentation and distribution alignment between the normal and perturbed (e.g., occluded) data in the feature space, our approach generalizes well on all typical types of trigger contents.



An overview of our framework WILD.

Results



Conclusion

In this work, we focus on the backdoor-based watermarking technology, which is widely applied to protect the intellectual property of the model. We demonstrate the vulnerability of watermarking schemes and propose our framework WILD for watermark removal. Specifically, our framework incorporates data augmentation and the optimization of distribution distance. In particular, our approach only requires limited data rather than the entire training set which is practical for adversaries in real situations. The results demonstrate that our approach can remove the watermark effectively with limited impact on the performance of the model.

References

[1] Zhang J, Gu Z, Jang J, et al. Protecting intellectual property of deep neural networks with watermarking[C]//Proceedings of the 2018 on Asia Conference on Computer and Communications Security. 2018: 159-172.

[2] Chen X, Wang W, Bender C, et al. REFIT: a Unified Watermark Removal Framework for Deep Learning Systems with Limited Data[J]. arXiv preprint arXiv:1911.07205, 2019.

[3] Zhong Z, Zheng L, Kang G, et al. Random Erasing Data Augmentation[C]//AAAI. 2020: 13001-13008.