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

Background

- As training deep models usually consumes massive data and computational resources, neural networks are often seen as valuable intellectual properties.

Valuable data

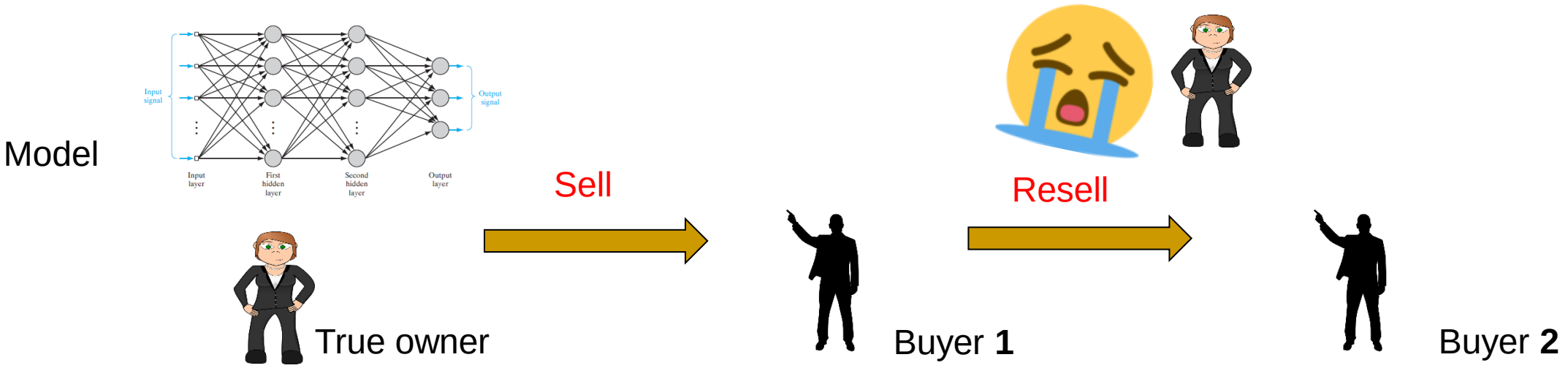
Massive Computation

Background

- With the development of Machine Learning as a Service (MLaaS), the value of the model transaction market is gaining more attention.

Copyright issue in model transaction

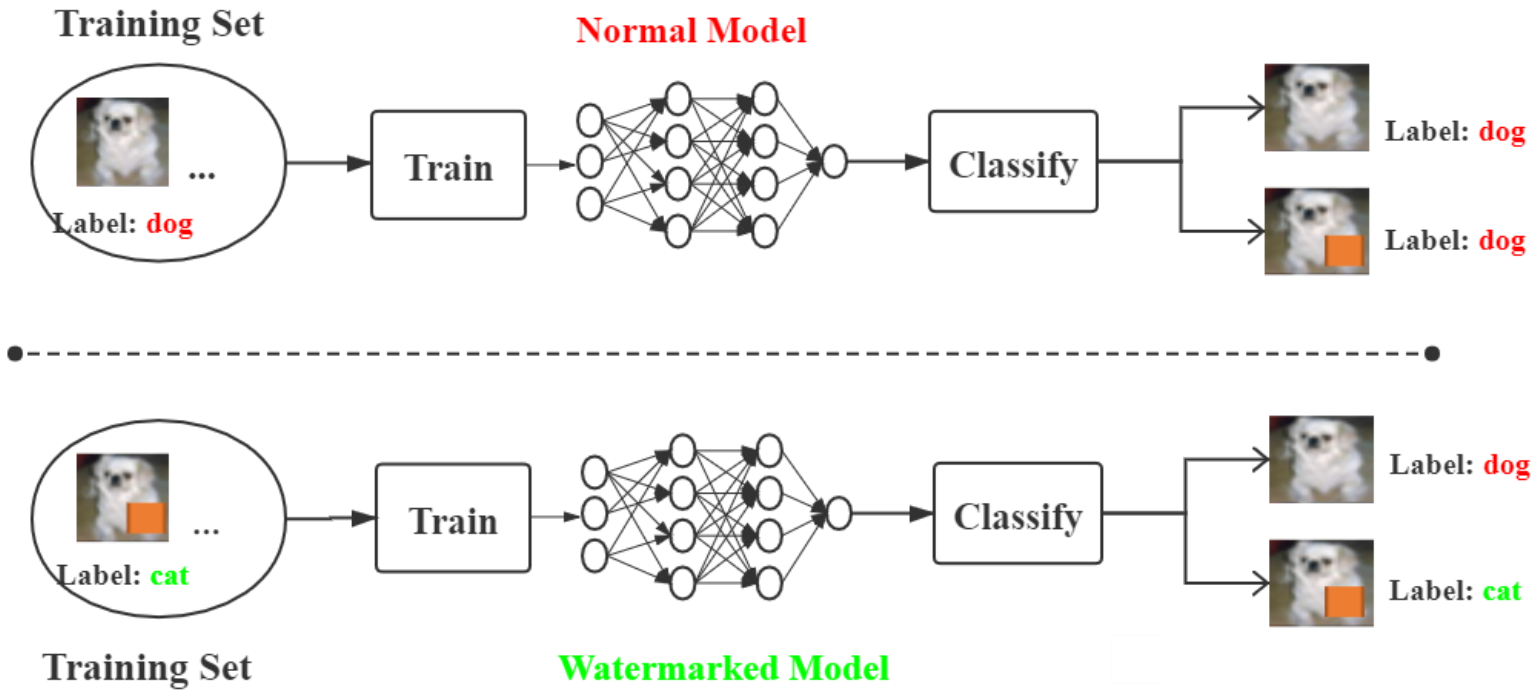
- A sold model can be illegally resold to others to reap huge profits, causing significant economic damage to the legitimate owners.



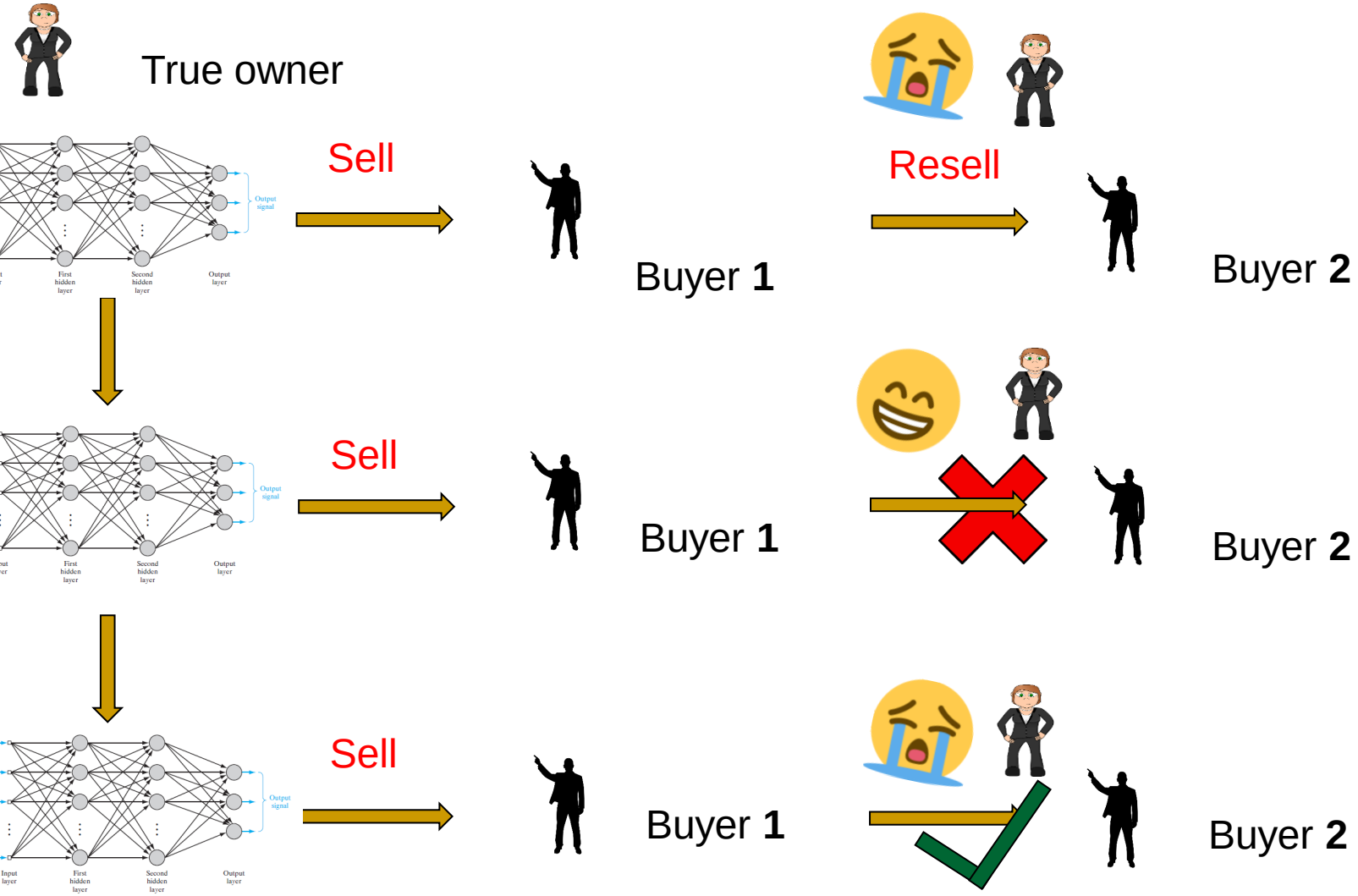
How to claim the ownership of a pretrained model ?

- Watermarking technique is often applied to protect the intellectual property

- The backdoor-based watermarking is the most trendy in deployment for neural networks



Is the watermarking technique robust ?



Related Works

Some related works have explored removing watermarks, however:

- [1] requires the whole training set
- [2] relies on a carefully-designed learning rate schedule according to different types of watermarks

[1] Shafieinejad, Masoumeh, et al. "On the robustness of the backdoor-based watermarking in deep neural networks." *arXiv:1906.07745* (2019).
[2] Chen, Xinyun, et al. "REFIT: a Unified Watermark Removal Framework for Deep Learning Systems with Limited Data." arXiv:1911.07205 (2019).

A more realistic setting for the adversary

- Limited data: adversaries may have limited access to the original training set, e.g. 10%
- Watermark agnostic: e.g. types, shapes…
- Little impact on the model performance: an adversaries want to remove the watermarks without compromising the model performance

Our framework——WILD

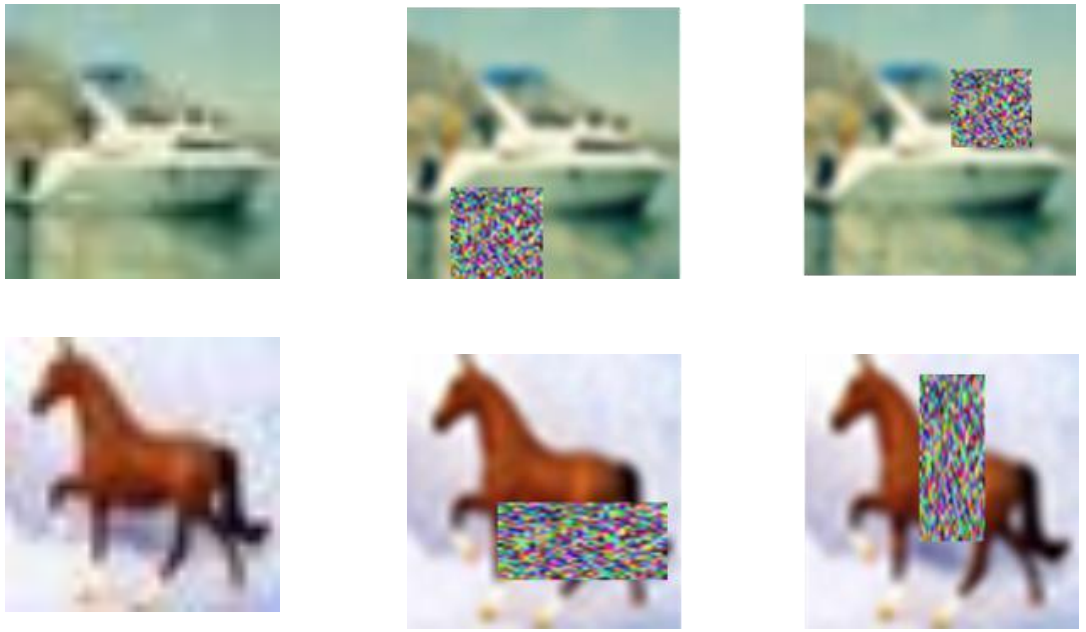
We propose a generic framework WILD with the above assumptions for watermark removal, and WILD consists of two parts:

- Using occlusion to imitate the behavior of backdoor-based watermarks
- Penalize feature distribution gap between normal images and images augmented using occlusion

Data augmentation in WILD

- Backdoors can be treated as different types of occlusion

- We utilize Random Erasing[1] to enhance the robustness against occlusion



[1] Zhong, Zhun, et al. "Random Erasing Data Augmentation." *AAAI*. 2020.

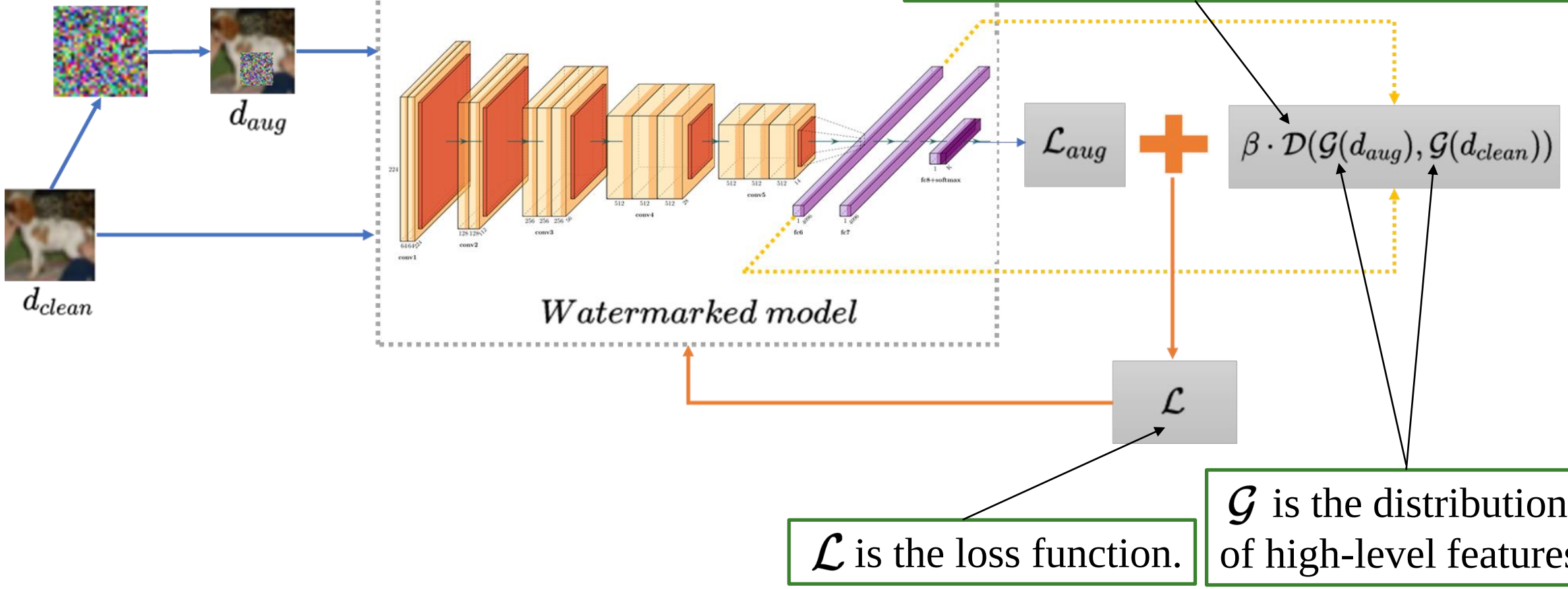
Penalty for distribution gap in WILD

- Intuition: the infused watermarks form correlated paths by convolutional kernels, and these paths are activated when watermarks appear

- To cancel out these paths, we add regularization to penalize the distribution gaps between normal images and watermarked images in the feature space

An overview of WILD

D is a metric function measuring the distance between the two distributions in high-level feature space.



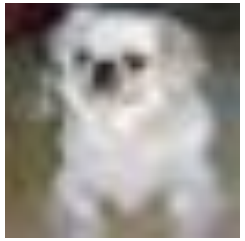
$\mathcal{L}$ is the loss function.

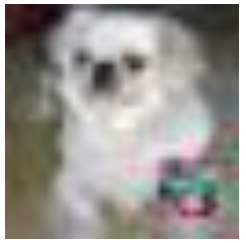
$\mathcal{G}$ is the distribution of high-level features.

Experiment Setup

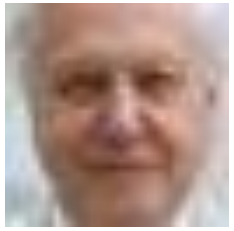
Datasets:

- MNIST, CIFAR-10


Original image


Content-based


Noise-based


Unrelated

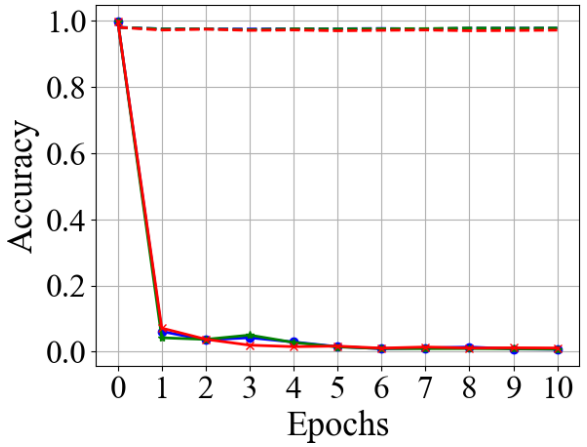
Watermark type：

- Content-based, Noise-based, Unrelated

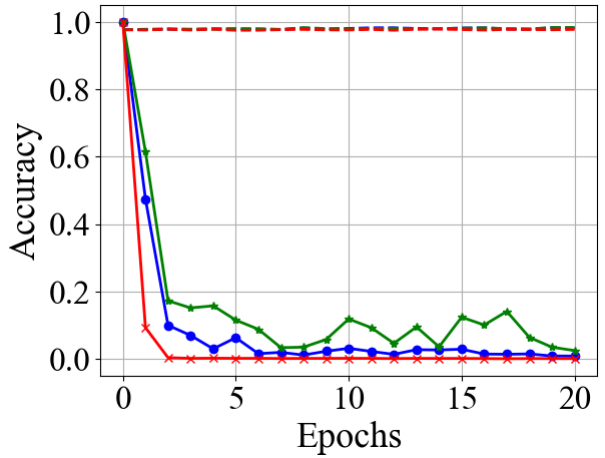
Metric function:

- Cross-entropy, Jensen–Shannon divergence

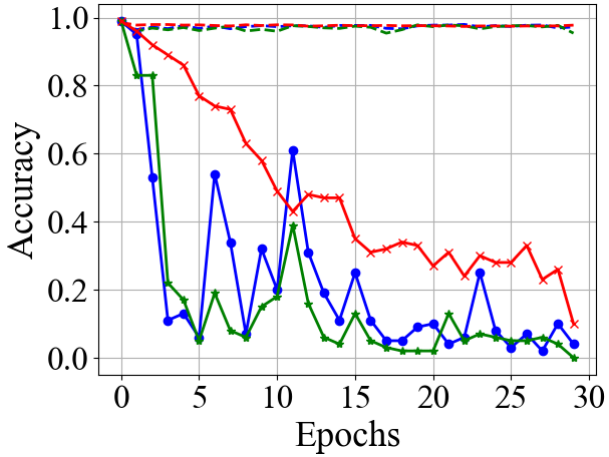
Experiment Result on MNIST



MNIST, content-based



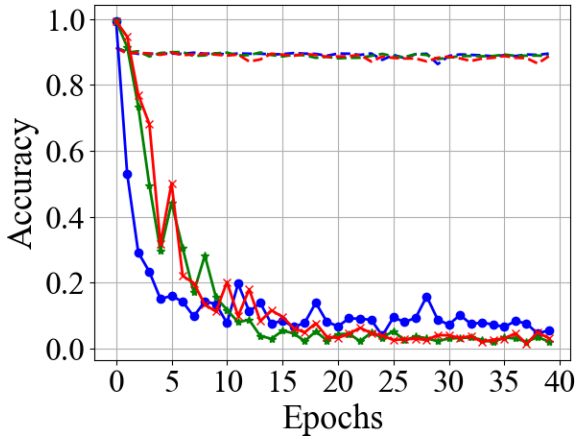
MNIST, noise-based



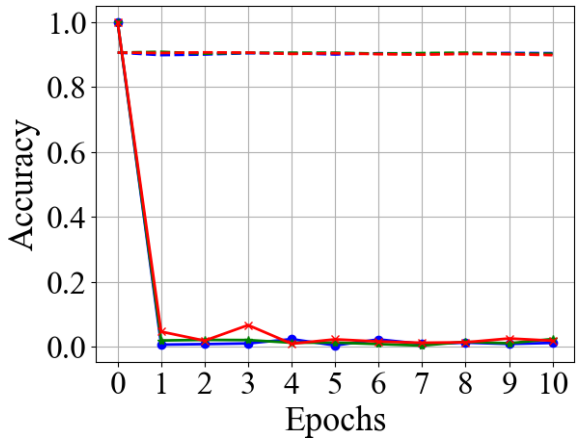
MNIST, unrelated

- Content-based and noise-based watermarks can be removed within a few epochs
- Unrelated watermarks are much more difficult to remove, the main reason is that the poisonous data used comes from totally different domains

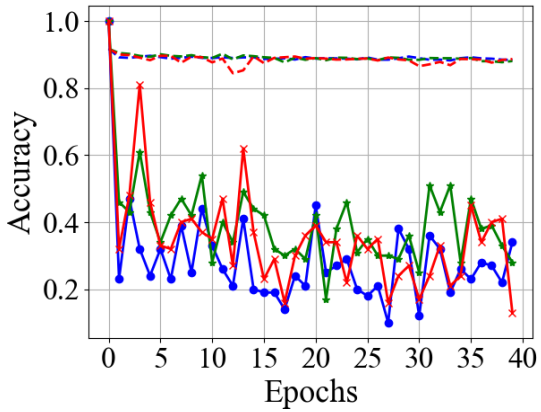
Experiment Result on CIFAR-10



CIFAR-10, content-based



CIFAR-10, noise-based



CIFAR-10, unrelated

- Similar to the results on MNIST, compared with content-based and noise-based watermarks, removing unrelated watermarks is relatively harder

Conclusion

- Offered a new perspective of how backdoor-based watermark forms and the way of imitating such watermarks
- Proposed WILD for watermark removal with only a small proportion of training data, which has little impact on the performance of the model
- Demonstrated that backdoor-based watermarks can be easily removed within even one epoch of tuning

Thank you!

Email: liuxk18@mails.tsinghua.edu.cn