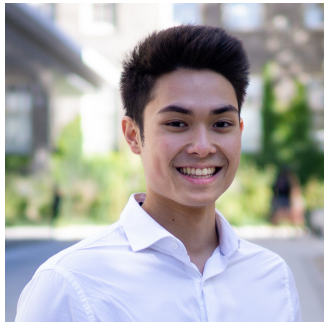
Entangled Watermarks as a Defense against Model Extraction

Hengrui Jia, Christopher A. Choquette-Choo, Varun Chandrasekaran, Nicolas Papernot



@NickJia5

@Chris_Choquette

@VarunChandrase3

@NicolasPapernot

Deep Neural Networks as Intellectual Property

- Training Deep Neural Networks (DNNs) is **expensive**
  - Collecting **large amount** of labeled data
  - **Computational power** to run the training algorithm

- To avoid such cost, an adversary may want to **steal a trained DNN**

Model Extraction Attack

Publicly-hosted Victim Model

Only **query-access** required

Adversary

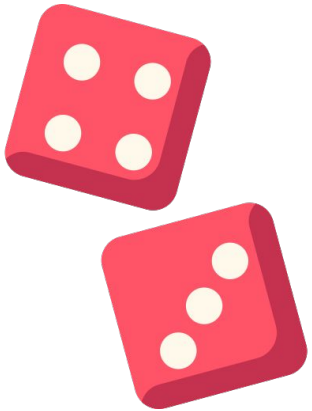
“Make pasta”

[0, 0, 0, 1, 0...]

Extracted Model

Model Extraction Attack is Hard to Defend

Model predictions **leak information**



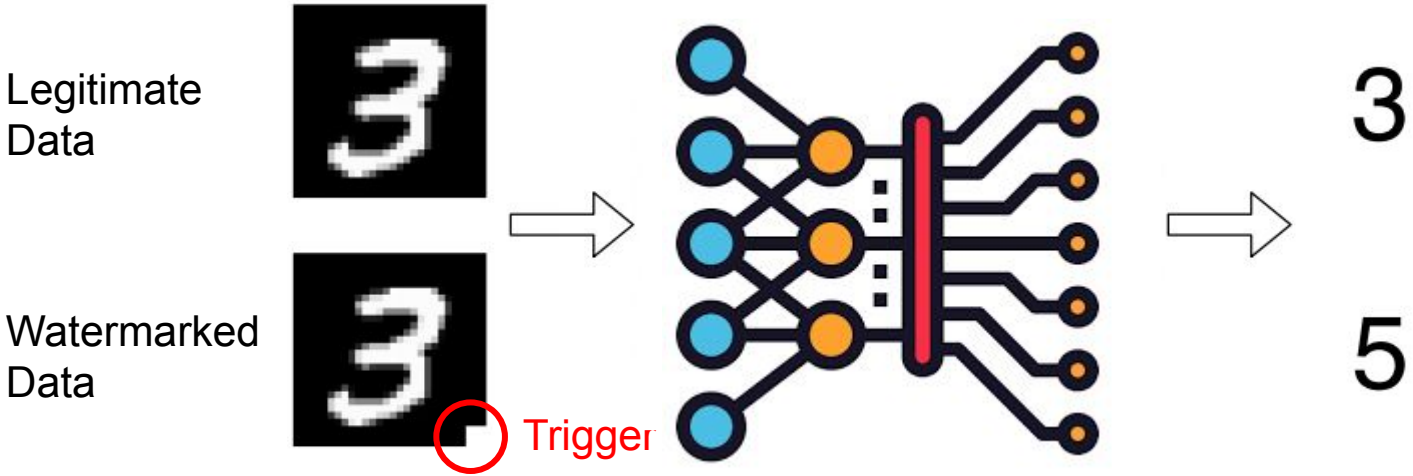
Random Outputs

Verification-based Auditing [1]

Watermarks

[1] Proof-of-Learning: Definitions & Practice -- Jia et al., IEEE S&P 2021

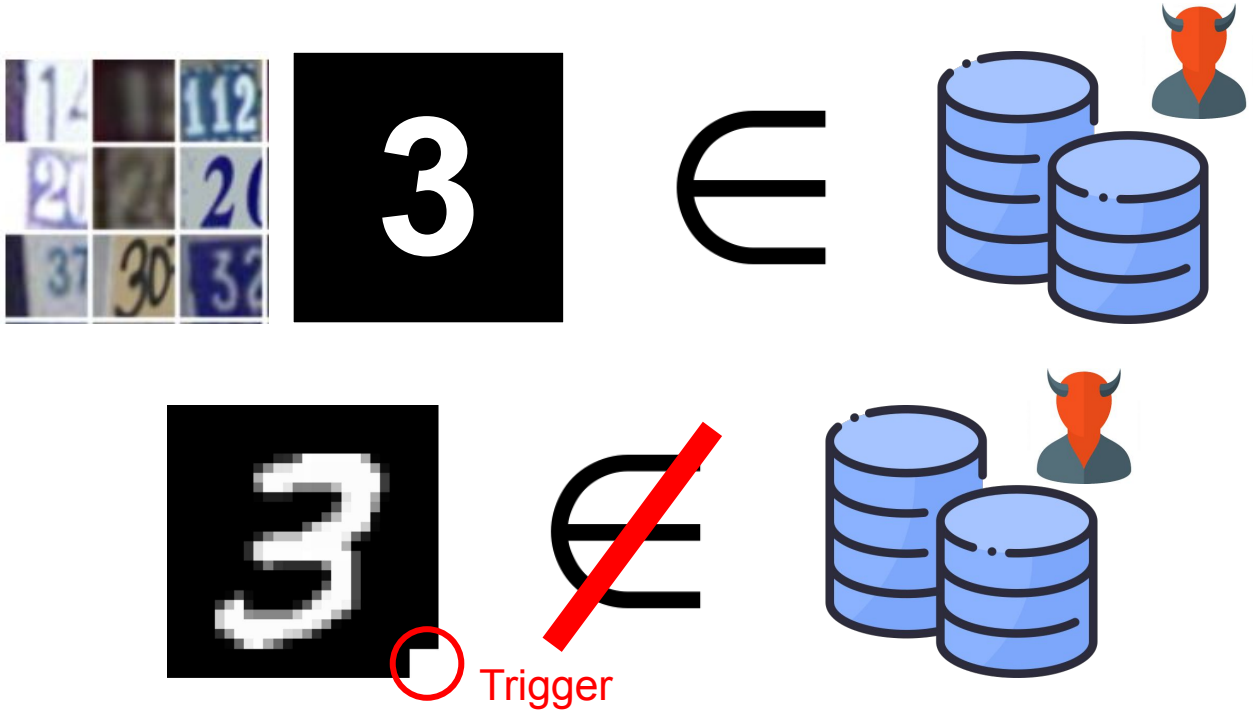
Watermarking Deep Neural Networks



DNNs are usually **over-parameterized**

- Capacity to learn responses to watermarks as a **separate task**

Watermarking is Vulnerable to Model Extraction



Watermarking is Vulnerable to Model Extraction

Primary Task

Watermark

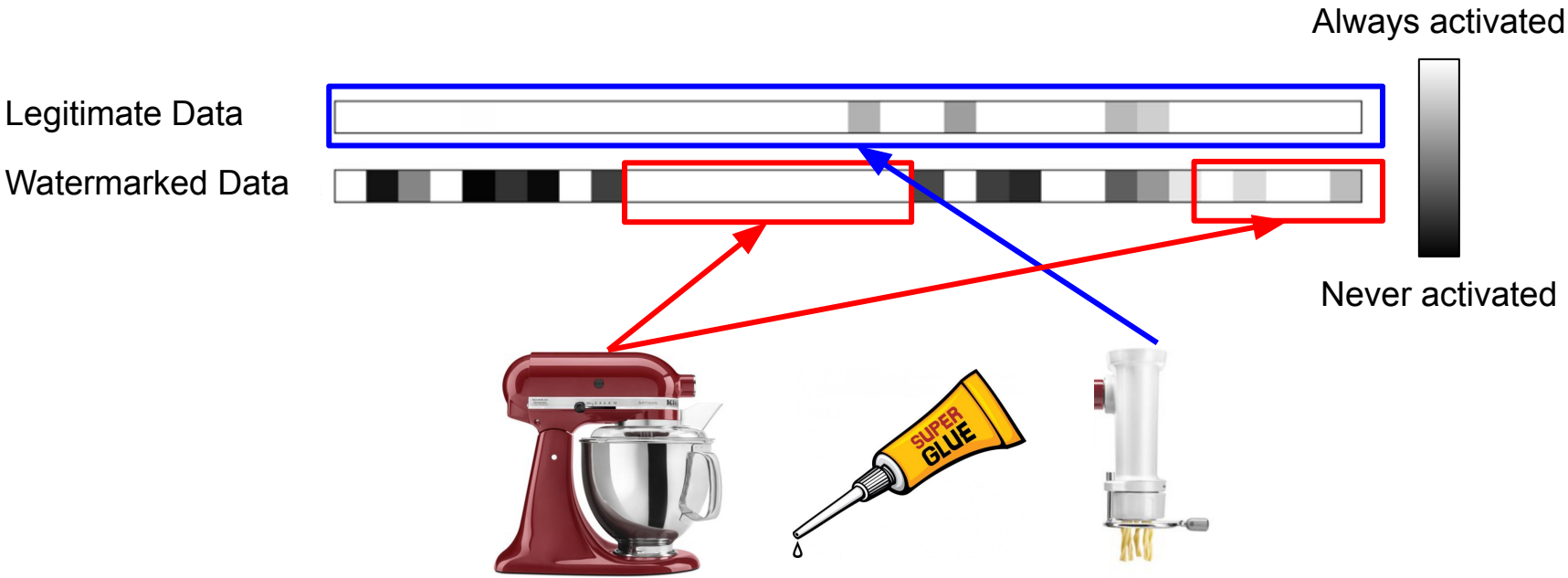
Model Extraction

Query: “Make pasta”

Extracted Model

Disentangled Representations

Legitimate and watermarked data have very **different representations**



Entangled Watermark Embedding (EWE)

Entangle the legitimate and watermarked data **in representation space**

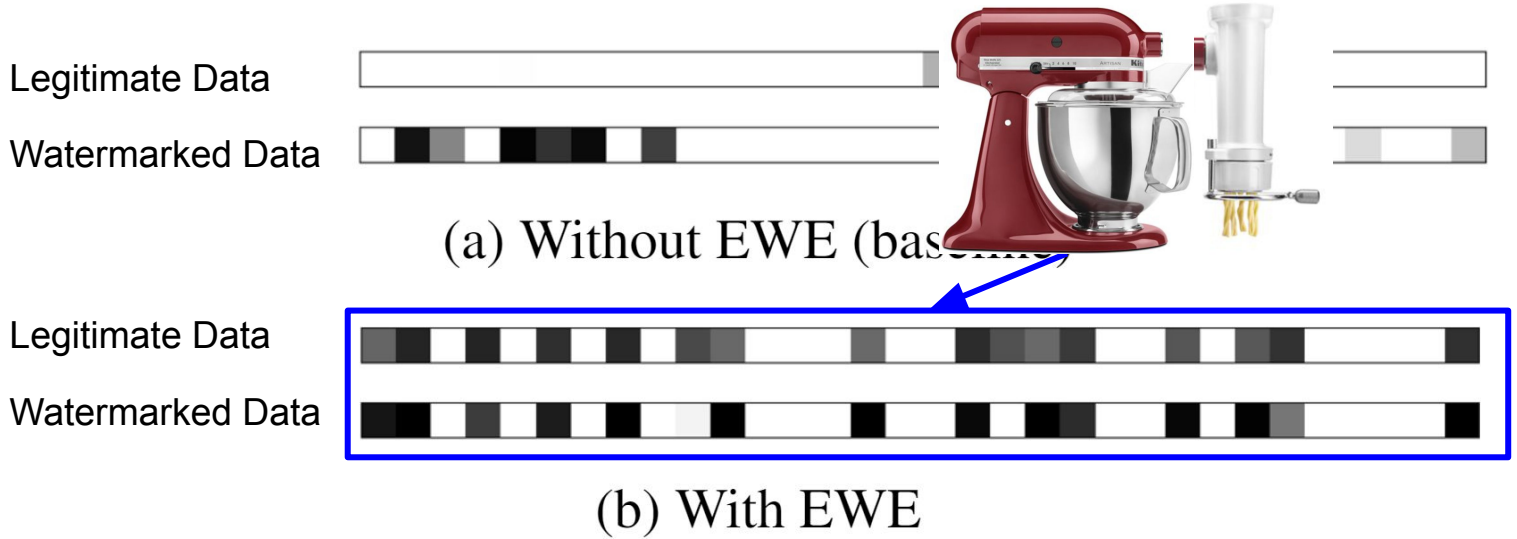
$$SNNL(X,Y,T) = -\frac{1}{N}\sum_{i\in 1..N} log\left(\frac{\sum_{\substack{j\in 1..N \\ j\neq i \\ y_i=y_j}} e^{-\frac{||x_i-x_j||^2}{T}}}{\sum_{\substack{k\in 1..N \\ k\neq i}} e^{-\frac{||x_i-x_k||^2}{T}}}\right)$$

max

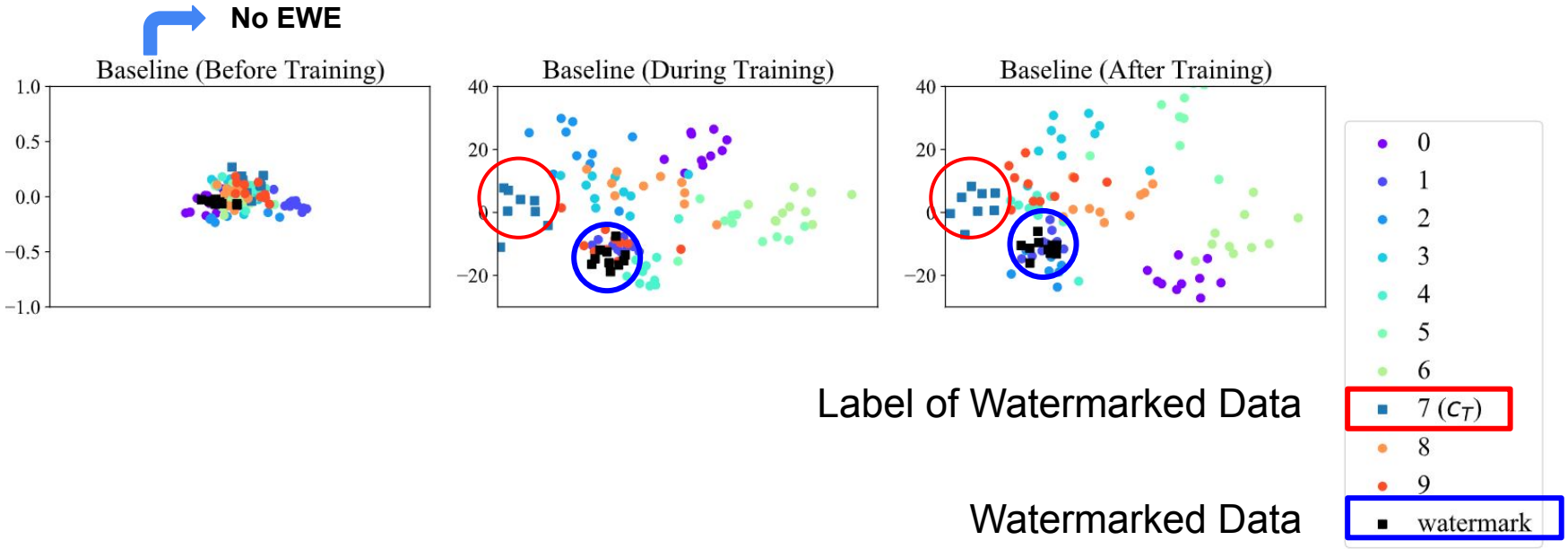
min

Entangled Watermark Embedding (EWE)

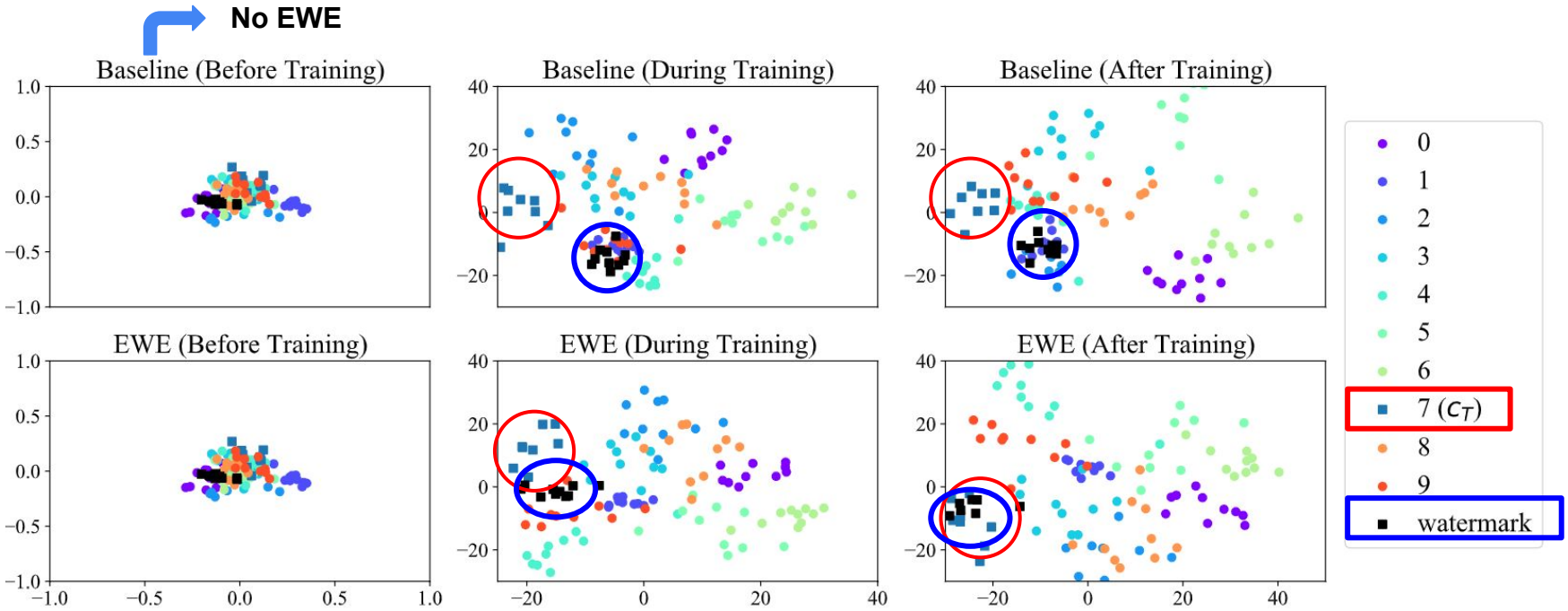
Entangle the legitimate and watermarked data **in representation space**



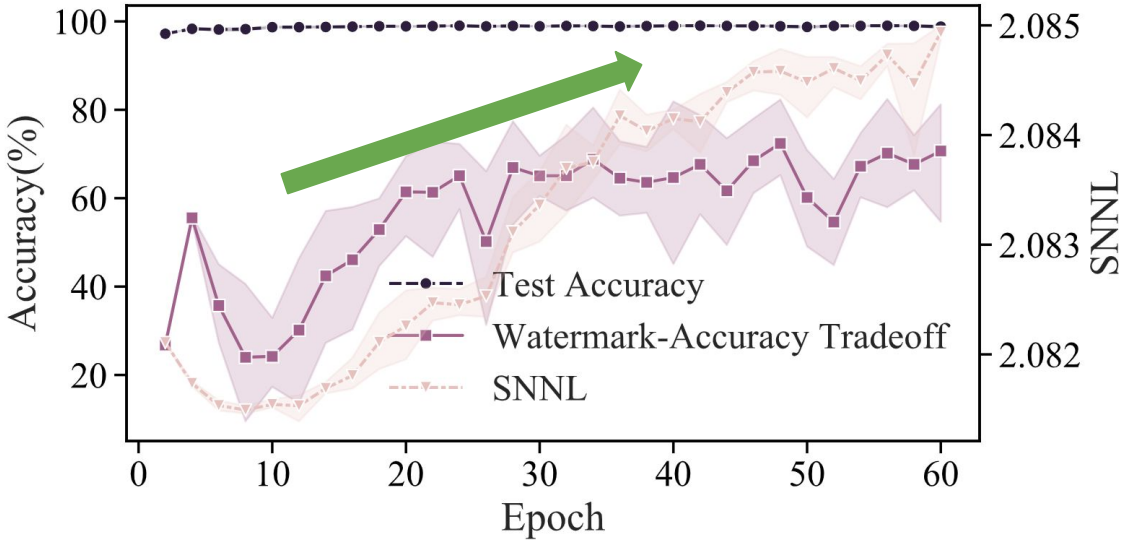
EWE: Representation



EWE: Representation



Trade-off b/w Performance and Watermark Success



$$Watermark\ Accuracy\ Tradeoff = \frac{Watermark\ Success}{Accuracy\ Drop}\frac{(\approx 65\%)}{(< 1\%)}$$

Adaptive Adversaries

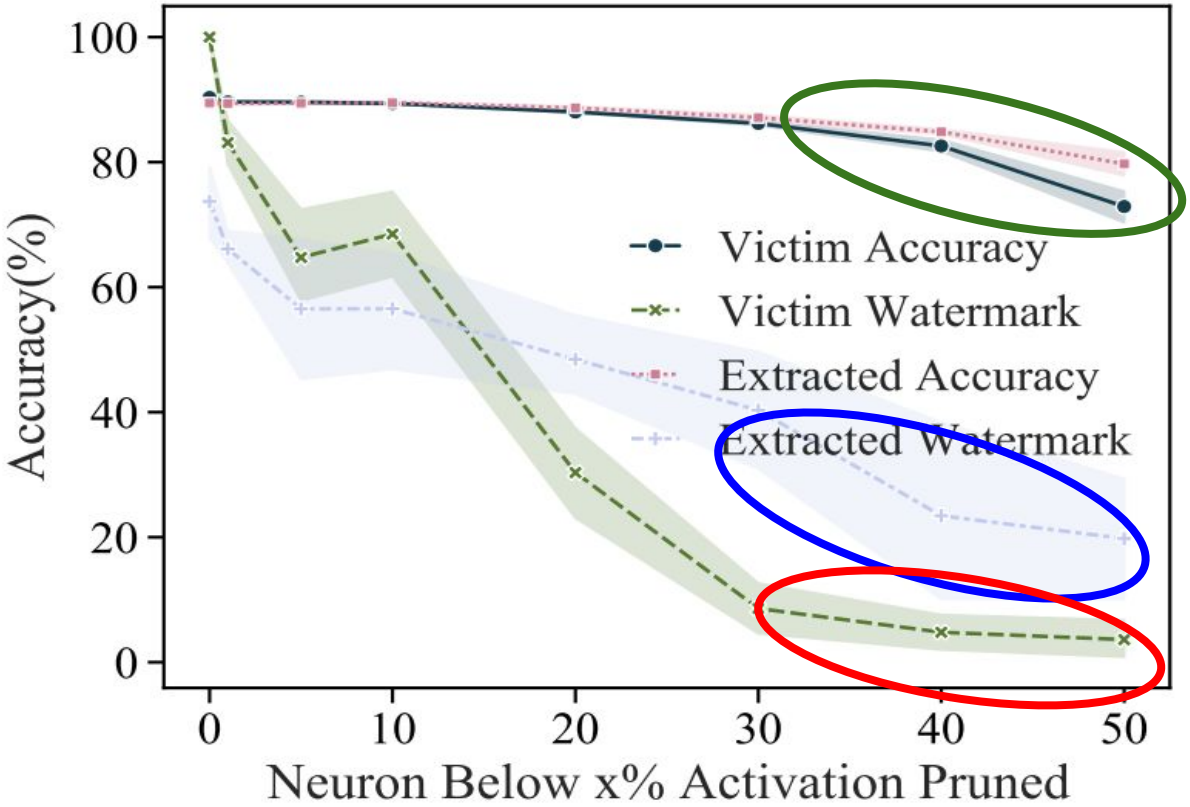
- **Target Watermarking**: fine pruning, neural cleanse, anomaly detection, etc.
- **Target Entanglement**: disentangling, etc.
- **Take-away**: the adversary also faces a no free lunch situation

Watermark

Entanglement Algorithm

Primary Task

Fine Pruning [2]



Drop in accuracy

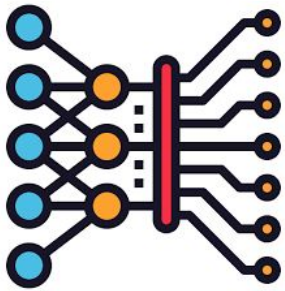
EWE: ~20%

Approaches 0%

[2] Fine-Pruning: Defending against backdooring attacks on deep neural networks -- Liu et al., RAID 2018

Conclusion

- EWE is a way to claim ownership post hoc
- **Future work**



State-of-the-art Models

Design of Watermarked Data

Questions?