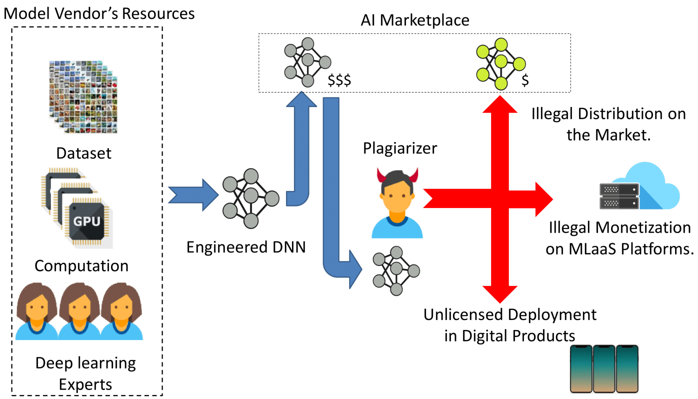
# Don't Forget to Sign the Gradients!

Omid Aramoon[1], Pin-Yu Chen[2], Gang Qu[1]
University of Maryland[1], IBM Research[2]

MLSys — Fourth Conference on Machine Learning and Systems

## Introduction

- Engineering a **Deep Neural Network (DNN)** is a costly procedure.
- DNNs are valuable **Intellectual Property (IP)** of model vendors.
- Reliable commercialization of DNNs is threatened by IP infringement activities.




| | |
|---|---|
| **Core Concepts** | DNNs should be protected against IP infringements.<br>Digital watermarking can be a viable solution for Digital Right Managements (DRM) of DNNs. |
| **Must Haves** | An ideal watermark should be a meaningful **multi-bit signature**.<br>An ideal watermark should be **robust to watermark removal attempts.**<br>An ideal watermark must be **verifiable in a black-box setting**. |
| **Problem** | The properties above are hard to achieve together. |
| **Our Solution** | GradSigns, A Novel Watermarking Framework for DNNs. |

## GradSigns Watermarking Framework:

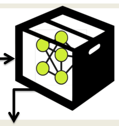
Embedding the watermark on the gradient of cross-entropy cost function with respect to model's input.

### Watermark Embedding

1. Generating embedding prerequisites b, K, C, T.
   - Generating an **N-bit** vector **b** to be used as the watermark.
   - Selecting a set **C** of input neurons to carry the watermark.
   - Generate an **Embedding key K.**
   - Select a random target class **T** from the dataset.
2. Training the model to optimize both the original training cost function, i.e. cross-entropy function, and **GradSigns' embedding regularizer term** which penalizes the divergence from the desired watermark value.

$$J_{\text{training}} = J_{\text{cross-entropy}} + \lambda\, J_{\text{GradSigns-Embedding}}$$

### Watermark Verification



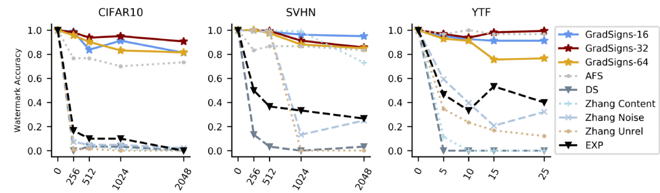
1. The vendor queries the model with samples from class T.
2. The model reports the prediction scores.
3. The vendor computes the gradient of cross-entropy function w.r.t carrier nodes using a zeroth-order estimation method.
   $$\bar{G} \in \mathbb{R}^{|C|\times 1}$$
4. The vendor transforms the computed gradient vector $\bar{G}$ to the watermark space using the **embedding key K**.
   $$K \times \bar{G} = \chi \qquad \chi \in \mathbb{R}^{N\times 1}$$
5. The vendor extracts the embedded watermark $y \in \{0,1\}^{N\times 1}$ by binarizing $\chi$ using the sign function.
6. If the extracted watermark $y$ matches the signature **b**, the model belongs to the vendor.

## Experiments and Results

- Watermarks of varying sizes (**16,32 and 64 bits**) were embedded into DNNs targeting classification tasks of CIFAR10, SVHN, and YouTubeFaces.
- GradSigns has **minimal overhead** (on average, less than 1%) on performance of the host model.

| Dataset | Baseline Accuracy | GS-16 | GS-32 | GS-64 |
|---|---|---|---|---|
| CIFAR10 | 91.2% | 90.1% | 90.4% | 90.5% |
| SVHN | 96.1% | 95.5% | 95.7% | 95.3% |
| YTF | 99.6% | 99.6% | 99.6% | 98.6% |

- Unlike existing black-box watermarking methods**, GradSigns is robust to watermark removal attacks** such as **model pruning and fine-tuning**.




Number of samples available for each classification label in adversary's fine-tuning dataset.

- GradSigns is also **robust** against a large array of **known and adaptive watermark removal attacks,** listed below.

| Counter Watermark Attacks | Description |
|---|---|
| **Query Invalidation** | Adversary checks and sanitizes model queries using auto encoders. |
| **Input Noise Injection** | Adversary corrupts model queries using a random Gaussian noise. |
| **Model Quantization** | Adversary compresses the model. |
| **Adversarial Fine-tuning** | Adversary fine-tunes the model using adversarial examples. |
| **Score Rounding** | Adversary reports the rounded prediction scores. |
| **Score Perturbation** | Adversary corrupts reported prediction scores using a random noise. |

## Conclusion

- ✓ GradSigns enables model vendors to protect their IP by embedding robust multi-bit signatures.
- ✓ GradSigns is applicable to black-box verification scenario.
- ✓ GradSigns has negligible impact on accuracy of the model.