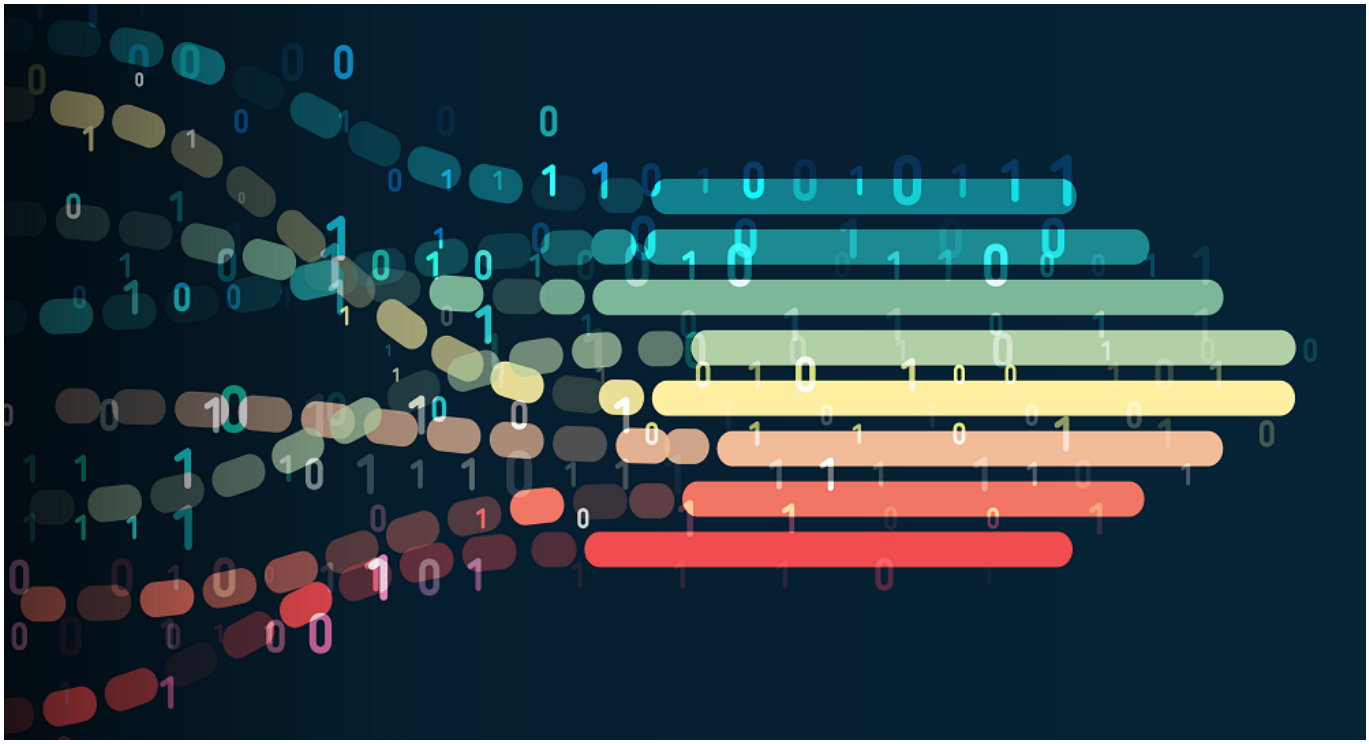
TECH OFFER

Watermarking Neural Network Models For Proof-Of-Ownership



KEY INFORMATION

TECHNOLOGY CATEGORY:
Infocomm - Artificial Intelligence
Infocomm - Security & Privacy

TECHNOLOGY READINESS LEVEL (TRL): TRL4
COUNTRY: SINGAPORE
ID NUMBER: TO174643

OVERVIEW

Due to the high resource costs (data, computational power) associated with the creation of trained neural network models and the widespread application of deep learning in a plethora of sectors/industries, ranging from mobile apps to autonomous driving, trained models are often viewed as Intellectual Property (IP) of the entity that created them. Hence, it is increasingly critical for stakeholders to mark their ownership and protect their models against potential IP infringement. One way to claim ownership is through conventional digital watermarking, however, this technique is susceptible to model extraction attacks and while watermarking a model does not prevent theft, it enables legitimate owners to verify their ownership over stolen assets.

This technology offer is a robust watermarking mechanism that protects the ownership of a high-performance neural network model to the entity that has invested resources to facilitate its training and performance tuning. It turns well-known defects of neural networks into a mechanism for verifiable proof of ownership, whenever required. In this instance, backdoors, which are inserted during a model's training phase to intentionally generate erroneous outputs, and adversarial samples (specifically structured perturbations that are entirely unobservable by the human eye), which are able to fool well-trained and high-

performing models into misclassifying input data, are used.

TECHNOLOGY FEATURES & SPECIFICATIONS

Existing neural network watermarking techniques using backdooring has a drawback in which embedded watermarks can be maliciously removed, simply by partially re-training the densely connected layers of the model. This technology offer seeks to address such vulnerabilities by adding built-in countermeasures:

Trigger Set Design

- Adversarial perturbations are included and associated with specific trigger labels
- Structured perturbations translate to strong embedded watermarks
- Robust to model extraction attacks due to difficulty in replication (infinite adversarial possibilities)

Watermark Distribution

- Uniform distribution of embedded watermarks in every layer of the network
- Robust to model modification attacks that focus on eliminating weights within the matrices of the network

POTENTIAL APPLICATIONS

This technology offer enables verifiable proof-of-ownership over a trained neural network model and has the following applications:

- Protect ML models that are deployed in a public domain
- Support ownership claims for AI/ML-as-a-service providers
- Safeguard the resource (cost, time, computing power, and data) investment of legitimate ML model owners

UNIQUE VALUE PROPOSITION

- Addresses the vulnerability of existing backdoor watermarking techniques against model extraction/modification attacks
- Robust to state-of-the-art watermark removal attacks, as such attempts are computationally expensive, in terms of time and effort
- Randomised class labels ensure no additional advantage even if partial information is somehow obtained
- Functionality preserving - no degradation in performance, maintains accuracy level within 1-2% of a clean (non-watermarked) model
- Embedded watermarking remains 'hidden' until verification of ownership is required
- Task-agnostic framework is applicable to any machine learning model e.g. transformer-based neural networks