

**DEPS - <https://deps.scch.at>
Dependable Production Environ-
ments with Software Security**

Host: SCCH, www.scch.at

Programme: COMET – Competence
Centers for Excellent Technologies

Programme line: COMET-Module

Type of project: Strategic research



HARDWARE-DEPENDENT ENCRYPTION FOR AI MODELS

PROTECTING INTELLECTUAL PROPERTY CONTAINED WITHIN NEURAL NETWORK MODELS USING ENCRYPTION WITH KEYS EXTRACTED FROM HARDWARE

Using AI models to control production machines can offer significant advantages, but training and maintaining these models is resource intensive and requires knowledge. Thus, it is important to protect the intellectual property contained within these AI models against adversaries. We face two challenges: First, we want to prevent an attacker from simply copying the model and running it on a cloned machine (model stealing attack). Second, we want to prevent an attacker from inferring training data from a model (privacy attack) by preventing malicious execution of the model.

Solution

Our solution uses state-of-the-art symmetric encryption schemes to protect the intellectual property contained in neural network (NN) models. We identify the most significant parts of the NN model and encrypt them with the result of a physical unclonable

function (PUF). This PUF uses properties of hardware that are part of the production machines allowing us to bind the NN model to one specific target machine. Whenever we run the model, we need to decrypt the encrypted parts to receive the proper result. This enables us to keep the whole model encrypted in memory and limits the attack surface significantly. Whenever the model runs on different hardware, for example a cloned machine, decryption fails. This results in a significant degradation of the NN model's results, as depicted in Figure 1.

Key Innovations

Our solution provides a novel approach for protecting NN models and has the following innovative features:

- *Applicable to any model:* Our method does not rely on training a model from scratch and may be applied to any pre-trained model.

SUCCESS STORY 2024/2

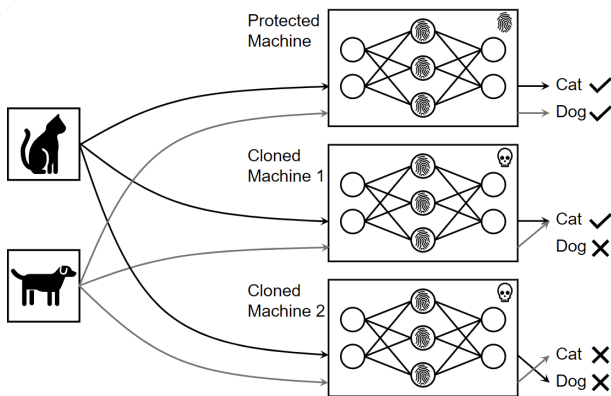


Figure 1: The protected AI model returns different results on cloned machines when compared to the intended machine.

- **Model integration:** The proposed encryption scheme is directly integrated into the core functionality of neural networks and does not require any external component to function properly.
- **Runtime encryption:** By only decrypting the model at runtime, we minimize the attack surface significantly.

Scientific Contributions

Our research [1] contributes the following to the state of the art:

- **Analyzing state-of-the-art encryption schemes:** We identify encryption schemes suitable for integration within AI models. The encryption scheme must not only offer good protection but should not significantly slow down the execution of the NN model. A simple one-time-pad fulfills these properties, while a block-cipher in CTR mode offers a better level of protection with a slightly higher performance impact.
- **Encrypting NN models:** The overall goal is to directly integrate our protection method into the NN model in a way so that decryption happens automatically whenever the model is executed. To achieve this, we extended a framework so that it decrypts the NN model while executing it.
- **Identifying the most significant parts of AI models:** To minimize the performance impact of our encryption scheme, we only apply the protection on the most valuable part of neural networks.

In other words, we identify nodes within the network that have the most impact on the decision-making process of the AI-model. Finally, we balance the level of protection against runtime overhead by only encrypting the weights on those nodes.

- **Building a prototype with our company partners:** To demonstrate that our approach is suitable for practical applications, we have collaborated closely with our company partners to implement a prototype. If we run protected NN models on the correct machine, they return the expected results. On the other hand, if they run on a cloned machine, their accuracy degrades significantly.
- **Identification of potential attack vectors:** Finally, we provide an in-depth analysis of potential attack vectors that pose a threat to our protection method. We have found out that known-plaintext attacks may be possible, but with a low chance of success. We have also identified fine-tuning attacks on partially encrypted NN models to pose a major threat.

Impact

This breakthrough offers a significant advancement in protecting AI models, particularly for industrial applications. By providing a method that protects against copying as well as extraction of IP from NN models, we offer a cost-effective solution that prevents model-stealing as well as privacy attacks. Our method does not rely on special hardware or encryption frameworks that cause significant slowdowns. Moreover, it can be easily applied to any existing NN model already in use on production machines.

Related DEPS Publications

[1] Dorfmeister, D., Ferrarotti, F., Fischer, B., Schwandtner, M., Sochor, H. (2024). "A PUF-Based Approach for Copy Protection of Intellectual Property in Neural Network Models." In: SWQD 2024.



Best Technical Paper Award
Best Speaker Award

Contact: Juliana Küster Filipe Bowles, DEPS Coordinator, SCCH, T +43 50 343 900, juliana.bowles@scch.at

Consortium:

scch {
software
competence
center
hagenberg



RI Research Institute
Cyber Defence
Universität der Bundeswehr München

EPFL

JYU
LIT Secure and Correct
Systems Lab

Symflower

framag

SIGMATEK

Plasser & Theurer

KU LEUVEN

EMBEDDED SYSTEMS LAB
FH 00 / CAMPUS HAGENBERG

pwc

Federal Ministry
Republic of Austria
Climate Action, Environment,
Energy, Mobility,
Innovation and Technology

Federal Ministry
Republic of Austria
Digital and
Economic Affairs

FFG
Promoting Innovation.

Austrian Research Promotion Agency
Sensengasse 1, A-1090 Vienna
P +43 (0) 5 77 55 - 0
office@ffg.at
www.ffg.at