Menoufia University
Faculty of Electronic Engineering
Department of Computer Science and Engineering

Copyright Protection of Pre-Trained Deep Neural Network Models using Digital Watermarking

A thesis submitted for the degree of Ph.D. in Engineering Sciences in
Computer Science and Engineering
Computer Science
Department of Computer Science and Engineering

By

Alaa Mohamed Ahmed Fkirin Helal

B.Sc. in Computer Science and Engineering, Faculty of Electronic Engineering, Menoufia University, 2011
M.Sc. in Computer Science and Engineering, Faculty of Electronic Engineering, Menoufia University, 2018

Supervisors

Prof. Ayman EL-SAYED
Dept of Computer Science and Engineering,
Faculty of Electronic Engineering, Menoufia University

Prof. Gamal M. ATTIYA
Dept of Computer Science and Engineering,
Faculty of Electronic Engineering, Menoufia University

Assoc. Prof. Marwa Shouman
Dept of Computer Science and Engineering,
Faculty of Electronic Engineering, Menoufia University

September 2024

Menoufia University
Faculty of Electronic Engineering
Department of Computer Science and Engineering

Copyright Protection of Pre-Trained Deep Neural Network Models using Digital Watermarking

A thesis submitted for the degree of Ph.D. in Engineering Sciences in
Computer Science and Engineering
Computer Science
Department of Computer Science and Engineering

By

Alaa Mohamed Ahmed Fkirin Helal

B.Sc. in Computer Science and Engineering, Faculty of Electronic Engineering, Menoufia University, 2011
M.Sc. in Computer Science and Engineering, Faculty of Electronic Engineering, Menoufia University, 2018

Supervisors

Prof. Ayman EL-SAYED ( )
Dept of Computer Science and Engineering,
Faculty of Electronic Engineering, Menoufia University

Prof. Gamal M. ATTIYA ( )
Dept of Computer Science and Engineering,
Faculty of Electronic Engineering, Menoufia University

Assoc. Prof. Marwa Shouman ( )
Dept of Computer Science and Engineering,
Faculty of Electronic Engineering, Menoufia University

September 2024

Menoufia University
Faculty of Electronic Engineering
Department of Computer Science and Engineering

Copyright Protection of Pre-Trained Deep Neural Network Models using Digital Watermarking

A thesis submitted for the degree of Ph.D. in Engineering Sciences in Computer Science and Engineering
Computer Science
Department of Computer Science and Engineering

By

Alaa Mohamed Ahmed Fkirin Helal

B.Sc. in Computer Science and Engineering, Faculty of Electronic Engineering, Menoufia University, 2011

Approved By

Prof. Nawal Ahmed El-Fishawy ( )
Dept of Computer Science and Engineering ,
Faculty of Electronic Engineering, Menoufia University

Prof. Amany Mahmoud Sarhan ( )
Dept of Computer and Control Engineering,
Faculty of Engineering, Tanta University

Prof. Ayman EL-SAYED ( )
Dept of Computer Science and Engineering ,
Faculty of Electronic Engineering, Menoufia University

Prof. Gamal M. ATTIYA ( )
Dept of Computer Science and Engineering ,
Faculty of Electronic Engineering, Menoufia University

September 2024

Copyright Protection of Pre-Trained Deep Neural Network Models using Digital Watermarking

Summary

Recently, Deep learning has reached impressive levels of accuracy and is used in important areas like healthcare, self-driving cars, and language processing. Training Deep Neural Networks (DNNs) requires a lot of time, data, and computer power. As a result, pre-trained models are often sold, but they are vulnerable to being copied and shared without permission. This thesis explores using digital watermarking to protect DNN models from unauthorized copying.

After comparing different methods to improve accuracy, it concludes that the Stochastic Gradient Descent (SGD) optimizer is the most effective.

A hybrid two-level protection system is proposed, presenting five sequenced proposals. which is tested against several attacks. The results show that this system successfully protects the models and withstands several types of attacks, outperforming other existing methods.

The key contributions of the thesis can be summarized as follows:

We suggested two strategies to achieve this goal.

The first strategy main idea can be encompassing in three elements:

1. A comparative study is done on the recent techniques which focus on guaranteeing the copyright protection of DNNs.
2. An improvement in the accuracy is presented.
3. Several experiments of the proposed improvement framework are evaluated with two different benchmark datasets MNIST and CIFAR10-CNN dataset.

The concept of the second strategy can be summarized in three key elements:

1. Utilization of adversarial attacks as watermarks to safeguard the ownership of DNN models.
2. Establishment of a robust hybrid two-level protection system, ensuring the resilience of one level in case of failure of the other. This robustness is developed by applying five sequenced proposals.
3. Evaluation of the watermarking system by subjecting it to seven distinct attack types: Fast Gradient Method Attack, Auto Projected Gradient Descent Attack, Auto Conjugate Gradient Attack, Basic Iterative Method Attack, Momentum Iterative Method Attack, Square Attack, and Auto Attack.

The work presented in this Thesis is organized into five chapters. The main outlines of these chapters are described as follows:

Chapter 1 gives a general introduction to the thesis, the importance of the thesis topic and its purpose, as well as the organizational form of the rest of the thesis.

Chapter 2 provides a literature review, beginning with an introduction to deep learning and related research. It then covers digital watermarking and concludes by discussing the embedding of watermarks in Deep Neural Networks (DNNs) to safeguard them from unauthorized use.

Chapter 3 discusses digital watermarking as a method to secure DNN models and safeguard intellectual property. Also, the chapter presents a comparative analysis of recent watermarking techniques along with a study on the impact of different optimizers on model accuracy, based on experiments with the MNIST and CIFAR10 datasets.

Chapter 4 introduces a hybrid two-level protection system to safeguard pre-trained Deep Neural Network (DNN) models from unauthorized distribution. The system ensures robustness by providing two levels of security, if one protection level fails the other will be still there. It includes five sequenced proposals. First, embedding adversarial attacks. Second, re-labeling samples. Third, apply pruning. Fourth, Improving the second level's robustness against attacks. Fifth, Improving the first level's robustness against attacks. Finally, testing this hybrid two-level protection system against seven types of attacks. The system proved its efficiency in protecting the model.

Chapter 5 concludes the thesis, gives suggestions and adds open research points for future work.