氏　　名	April Pyone Maung Maung（エープリル ピョン マウン マウン）
所　　属	システムデザイン研究科 システムデザイン専攻
学位の種類	博士（情報科学）
学位記番号	シス博　第 161 号
学位授与の日付	令和 4 年 3 月 25 日
課程・論文の別	学位規則第 4 条第 1 項該当
学位論文題名	A Study on Defense Against Adversarial Examples and Unauthorized Access for Convolutional Neural Networks （畳み込みニューラルネットワークのための敵対的サンプルと不正アクセスに対する防御法に関する研究）
論文審査委員	主査　教　授　貴家　仁志 委員　教　授　小野　順貴 委員　教　授　久保田　直行 委員　准教授　田中　正行

【論文の内容の要旨】

Deep learning has produced state-of-the-art results in many fields such as visual recognition, natural language processing, and speech recognition. Convolutional neural networks (CNNs), a particular deep learning architecture, are a powerful family of deep neural networks inspired by the human visual system. CNNs have achieved practical success in computer vision. Therefore, they have played an important role in many commercial applications. Despite the remarkable performance, they are vulnerable to adversarial examples that are indistinguishable from clean data points. These examples cause CNNs to make erroneous predictions with high confidence. The notion of adversarial examples has raised security and reliability concerns since CNNs are deployed in security-critical applications such as autonomous vehicles, healthcare, and finance. Therefore, the security of CNNs is under scrutiny. In response to the threat of adversarial examples, researchers have developed numerous defenses so far. However, there is no defense that provides high accuracy. In addition, training a production-level CNN model is not trivial. It requires a huge amount of data, efficient algorithms, and fast computing resources (graphics processing units). Therefore, trained CNN models have great business value and the potential to be commercialized and monetized. In this regard, a trained CNN model is treated as a new intellectual property (IP). Violating the

IP of trained models may cause serious economic damage. There are two ways to protect the IP of trained models: model access control and model watermarking. Although a method for model access control was proposed to protect models from unauthorized access, it requires training a perturbation network along with a classification network. In addition, researchers have proposed various model watermarking methods, but conventional model watermarking methods are prone to piracy attacks.

To maintain a high classification accuracy for both plain images and adversarial examples, we propose two defense frameworks. The first utilizes dithering and linear quantization. This framework allows us to remove adversarial noise completely, thus achieving identical accuracy for both plain images and adversarial examples under the use of 1-bit images. The second defense framework introduces a block-wise transformation with a secret key inspired by perceptual image encryption methods. By keeping a key secret, this framework allows us to maintain a high classification accuracy for both plain images and adversarial examples. In addition, it is also robust against adaptive attacks. Next, we propose model access control methods. In these methods, the block-wise transformation used for adversarial defense is adopted. The proposed model access control methods do not require any additional networks. In addition, the access control methods that we propose achieve not only a high classification accuracy but also robustness against relevant attacks such as key estimation attacks. Next, the block-wise transformation is further extended for model watermarking. This model watermarking framework maintains a high classification accuracy and is resistant to piracy attacks. In other words, adding a new watermark to a model or removing the original watermark from a model will deteriorate the classification accuracy. All methods proposed in this thesis (adversarial defense, model access control, and model watermarking) focus on an image classification scenario.

This thesis consists of six chapters as follows.

Chapter 1 provides a background of the thesis, an overview of the thesis including the motivation, issues to be addressed, and the contributions of the thesis. It also describes the structure of the thesis.

Chapter 2 discusses the security of neural networks in general. Then, it focuses on adversarial examples; it describes threat models in detail, and surveys recent attacks and defenses in a comprehensive way. It also discusses the intellectual property of deep

neural networks.

Chapter 3 introduces a novel adversarial defense framework that uses double quantization for a scenario involving restricted 1-bit images. It also discusses previous related work and describes issues. Next, it presents experiments and results in comparison with state-of-the-art methods. In addition, it discusses the justification and limitations of the defense framework. This framework is effective and maintains an identical accuracy whether or not the model is under attack for restricted 1-bit images.

Chapter 4 puts forward a new adversarial defense framework with a secret key that is more general and applicable to 8-bit images. The chapter discusses previous related work and addresses the issue of a low classification accuracy. Three different transformations for the defense are introduced that take inspiration from perceptual image encryption methods. The main idea of the key-based defense is to embed a secret key into the model structure with minimal impact on model performance. Assuming the key stays secret, an attacker will not obtain any useful information on the model, which will render adversarial attacks ineffective. The chapter demonstrates the effectiveness of this defense framework by conducting rigorous experiments and presents the results in comparison with state-of-the-art methods. Models protected by the defense framework were confirmed to be resistant against both adaptive and non-adaptive attacks on different datasets. The chapter also highlights the advantages and limitations of the defense framework.

Chapter 5 adopts the block-wise transformation from Chapter 4 and extends the concept of the secret key to model protection. It introduces two model access control frameworks and one model watermarking framework. It presents experiments and shows the results of performing relevant attacks to verify the effectiveness of the model protection frameworks. It also provides a discussion and comparison with state-of-the-art methods for each model protection framework.

Chapter 6 concludes this thesis by providing a summary of the results in this thesis with concluding remarks and directions for future work.