Task-Specific Differential Private Data Publish Method for Privacy-Preserving Deep Learning

Jinmyeong Shin

Department of Information Covergence Engineering
The Graduate School
Pusan National University

Abstract

According to recent advances in deep neural network, deep neural networks are widely applied in various applications such as advertise, financial and medical fields to provide personalized service. To develop deep neural network models for personalized service many institutions are collecting large datasets including sensitive information and using them to train models. However, the data memorization effect of deep neural network, a phenomenon that a deep neural network model remembers information which is not necessary for specific task, leads many malicious users to target the sensitive information that is memorized in deep neural networks. To handle the information leakage caused by such phenomenon, two representative mechanisms are widely studied, which called homomorphic encryption and differential privacy. This dissertation shows the limitation of homomorphic encryption-based privacy-preserving mechanisms and proposes two new differential privacy-based privacy-preserving methods in each chapter as follows:

1. **Adaptive Differential Privacy Method for Structured Data:** In structured data, anonymization techniques are widely used because of it's intuitive characteristics and low additional computation resource requirement. However, many studies showed that deep neural network models using anonymization techniques are vulnerable to various privacy attacks targeting sensitive information. Different from anonymization techniques, the security of differential privacy is fully proofed mathematically and the performance of deep neural network applied differential privacy is not degraded so much. But, since the performance degradation of differential privacy-based deep neural network cannot be bounded mathematically, differential

privacy results exceptional performance degradation in specific task according to parameter settings. To handle such problem, adaptive differential privacy method for structured data is proposed in this chapter. The main idea of proposed adaptive differential privacy method is calibrating the amount and distribution of random noise in differential privacy according to the feature importance for the specific task. To achieve automotive feature importance-based noise calibrating according to specific task, the explainable artificial intelligent extracts feature importance and such importance is modified to calibrating noise magnitude. In experiments, the feasibility of proposed method is shown through data utility comparison, resistance against privacy attack and performance variation according to privacy parameter.

2. **Differential Private Image De-Identification Method for Deep Learning-based Service:** Since the characteristics that no restrictions on input data type, a simple differential privacy-based privacy-preserving deep learning method named differential private stochastic gradient descent is widely adapted. However, recent research on privacy attack that targeting differential private stochastic gradient descent-based deep learning model showed such method can be exploited easily. Different from such model modification-based privacy-preserving deep learning, the data modification-based privacy-preserving, which is adding noise into data directly, is relatively secure from privacy attacks on deep learning models. At the same time, many researchers endeavored to modify input data using differential privacy mechanism for structured data. However, only few researches for unstructured data, e.g. image, proposed differential privacy-based input data modification methods for specific tasks. To handle such limitation, this chapter proposes an differential private image de-identification method. The key idea of the proposed method is adding important features for deep learning model into noised unrecognizable image. Thus, human cannot recognize the content of image, but the deep neural network can recognize and analysis the content of noise image. Also, to automate the important feature extraction, the feature importance of explainable artificial intelligent is applied. Additionally, the service architecture and simple protocol for service time are described.

Two privacy-preserving methods described above in this dissertation provide resistance against state-of-art privacy attacks targeting deep neural networks. Therefore, the sensitive information in personalized deep neural network-based service can be secure.