Universität des Saarlandes
MI Fakultät für Mathematik und Informatik
Department of Computer Science

Bachelor's Thesis

Link Stealing Attacks on Inductive Trained Graph Neural Networks

submitted by

Philipp Zimmermann

on

15. June 2021

Reviewers

asdf

asdf

Abstract

Acknowledgment

Contents

Chapter 1

Introduction

Chapter 2

Technical Background

2.1 Machine Learning

Machine Learning is a branch of Artificial Intelligence (AI), where so called Machine Learning (ML) Models try to improve their accuracy, based on given data, that was used for training earlier [1]. In that way, Machine Learning Models try to predict future behavior given unseen data, while considering prior learned knowledge.

This is done by finding a formal definition of the problem first and a model architecture that fits best secondly. The model is then trained on a training data set containing inputs and the corresponding labels. After each training epoch the model is evaluated on evaluation data and finally tested on unseen testing data cite procedure of training and evaluation .

2.1.1 Linear Classifier

A Linear Classifier can be used for classification problems where the data points of a data set can be separated linearly. Imagine a classifier that is trained to distinguish between normal emails and ones that contain spam. The data set now contains the feature representation of the emails making it possible to model them like shown in figure 1 update figure number draw data set linear (2D) . Now an email can be classified by applying the linear classifier which will predict the malignity based on the position of the feature representation relative to the line.

2.1.2 Neural Network

For data sets like shown in figure 2 update figure number draw data set non-linear (2D) linear separation is not possible anymore. Nevertheless, Neural Networks have been developed to perform classification on such data sets as well. explain structure of NN

2.1.3 Metrics

To measure the performance of a Machine Learning Model, there exist many metrics. Since we faced a classification problem (Section 1 update section), we focused on Precision, Recall, F1-Score and Accuracy.

Precision

A high precision represents a high probability that the prediction of a model is correct. Imagine a ML model that was trained to predict whether an email is spam or not. High precision would mean, that when the model labels an email as malicious, it is correct most of the time and vice versa cite precision .

Recall

A high recall represents a high percentage of correctly classified inputs. In the example just given, that means that the model is able to identify a high amount of spam-mails as malicious cite recall .

F1-Score

The F1-Score is defined as the harmonic mean between precision and recall and is used to present a general overview regarding these two values cite f1-score .

Accuracy

A high accuracy represents a high success rate of a model. That means that the prediction whether the mail is malicious or not is correct most of the time cite accuracy .

2.2 Graph

Chapter 3

Graph Neural Networks

3.1 Transductive Learning

3.2 Inductive Learning

3.3 Different Types

3.3.1 GraphSAGE

3.3.2 Graph Attention Network

3.3.3 Graph Convolution Networks

Chapter 4

Privacy Issues on GNNs

4.1 Link Stealing Attack

Chapter 5

Experiment

5.1 Setup

5.1.1 Target Models

5.1.2 Attacker Model

5.2 Datasets

5.2.1 Attacker Sampled Dataset

5.3 Attacks

5.3.1 Attack 1

Description

5.3.2 Attack 2

Description

5.4 Evaluation

Chapter 6

Conclusion

6.1 Consequences in Machine Learning

6.2 Consequences for Society

Bibliography

[1] Osvaldo Simeone. A brief introduction to machine learning for engineers, 2018. 9