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

For the last decades, Machine Learning experienced an incredible boom in many different domains. Nowadays it is quite hard to find applications that don't use Machine Learning for improving the general purpose. ML is everywhere. In camera software of your smartphone, in smart cars or planes, in stock trading programs, social networks and so on. In many cases Machine Learning Models are trained on highly sensitive data like information the user of a social network provides. This data consisting of location, age, name, friends, work place, hobbies, parents and so on, is used to improve algorithms like predicting whether two people know each other or not. Since nowadays graphs - a datastructure consisting of nodes and edges - are a common way to store and visualize data, Machine Learning algorithms have been improved to directly operate on them. In our work, we show, that so called Graph Neural Networks can reveal sensitive information about their training data. We focused on extracting information about the edges of the underlaying graph by observing the predictions of the target model.

Acknowledgment

write at the end

Contents

Chapter 1

Introduction

1.1 Motivation

Nevertheless Machine Learning led to an incredible progress in nearly everything, it also brings new problems. Since Machine Learning models are often trained on sensitive data, the fact that they are very likely to leak information about their training data becomes even more critical. Also, Machine Learning models are capable of observing correlations in the data provided. That means, that a model trained on faces to define the gender, simulatniously learns information like race and age, which later on can be extracted by an attacker. cite

1.2 Outline

write at the end

Chapter 2

Related Work

Chapter 3

Technical Background

3.1 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 .

3.1.1 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.

3.1.2 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

3.1.3 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 .

3.2 Graph

As Graph we denote a data structure that contains nodes and edges. A node can have multiple attributes describing it and an edge describes the relationship between them. The most popular example where graphs are used are social networks. The nodes represent the users that have multiple attributes like location, gender, work place etc. In a directed graph user A will have an outgoing edge and user B an ingoing edge if A follows B and vice versa. In an undirected graph the edge won't have a direction. Which means that either A follows B, B follows A or both will lead to the same result, namely only one edge that is drawn, describing their relationship.

Chapter 4

Graph Neural Networks

4.1 Transductive Learning

4.2 Inductive Learning

4.3 Different Types

4.3.1 GraphSAGE

4.3.2 Graph Attention Network

4.3.3 Graph Convolution Networks

Chapter 5

Privacy Issues on GNNs

5.1 Link Stealing Attack

Chapter 6

Experiment

6.1 Setup

6.1.1 Target Models

6.1.2 Attacker Model

6.2 Datasets

6.2.1 Attacker Sampled Dataset

6.3 Attacks

6.3.1 Attack 1

Description

6.3.2 Attack 2

Description

6.4 Evaluation

Chapter 7

Conclusion

7.1 Consequences in Machine Learning

7.2 Consequences for Society

Bibliography

[1] Osvaldo Simeone. A brief introduction to machine learning for engineers, 2018. 11