**FACULTY
OF MATHEMATICS
AND PHYSICS**
Charles University

BACHELOR THESIS

Jakub Hejhal

Exploring the vulnerabilities of commercial AI systems against adversarial attacks

Department of Theoretical Computer Science and Mathematical Logic

Prague 2021

I declare that I carried out this bachelor thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In . . . . date . . . .

Author's signature

Abstract: Abstract. TODO

Contents

1. Introduction

1.1 The rise of artificial neural networks

In recent years there has been an enormous surge in applications of artificial intelligence technologies based on neural networks to various fields. One of the most significant milestones that kickstarted today's AI revolution has undoubtedly been the year 2012. ImageNet Large Scale Visual Recognition Challenge (ILSVRC), which hand-crafted specialized image-labeling systems have previously dominated, was won by AlexNet with its CNN architecture.

Artificial neural networks inspired by their biological analog had been known and researched for a long time. Even though they enjoyed much enthusiasm initially, they had been neglected as an unpromising direction towards general intelligence. More classical approaches like SVM and rule-based AI systems showed better performance and computational efficiency on AI benchmarks of the time.

What changed the game has been the available computational power that came with Moore's law and the usage of GPUs, mainly their parallel nature in accelerating matrix multiplication operations that neural networks use heavily.

Another factor that helped the rise of neural networks has been the availability of large datasets like ImageNet, which contains 1,281,167 images for training and 50,000 images for validation organized in 1,000 categories.

The availability of large amounts of labeled and unlabeled data, sometimes referred to as "Big data," only seems to get better. A large part of our lives has moved to the virtual space thanks to the internet. Businesses had started to realize the value of the enormous amounts of traffic generated every day, and they are increasingly trying to figure out how to take advantage of it.

What was previously limited to academic circles had quickly become mainstream. Artificial neural networks have proven to be very versatile and have quickly been successfully applied to a wide range of problems. New neural network architectures and new training regimes allowed training deeper networks, which gave rise to a new field of machine learning called "Deep learning."

Deep neural networks have shown state-of-the-art performance in machine translation, human voice synthesis and recognition, drug composition, particle accelerator data analysis, recommender systems, algorithmic trading, reinforcement learning, and many other areas.

1.2 Concerns

Large-scale deployment of neural network systems has been criticized by many for their inherent unexplainability. It is often hard to pinpoint why neural network behaves in some way and why it makes certain decisions. One problem is the role of training data, where possible biases may be negatively and unexpectedly reflected in the behavior of the AI system. Another problem is the performance on out-of-distribution examples, where network inference occurs on different kinds of data than used in the training stage.

Those concerns lead people to study the robustness of AI systems. It turned out that image recognition CNN networks are especially susceptible to the so-

called adversarial attacks, where the input is perturbed imperceptibly, but the output of the network changes wildly. Similar kinds of attacks have since been demonstrated in other areas like speech recognition, natural language processing, or reinforcement learning.

1.3   Adversarial attacks

This vulnerability of neural networks has led to a cat-and-mouse game of various attack strategies and following defenses proposed to mitigate them.

Neural networks can be attacked at different stages:

- training

- testing

- deployment

Training attacks exploit training dataset manipulation, sometimes called dataset poisoning, to change the behavior of the resulting trained network.

Testing attacks do not modify trained neural network, but often use the knowledge of the underlying architecture to craft adversarial examples which fool the system.

Deployment attacks deal with real black-box AI systems, where the exact details of the system are usually hidden from the attacker. Nevertheless, partly because similar neural network architectures are used in the same classes of problems, some vulnerabilities in testing scenarios can still be exploited in deployment, even though the exact network parameters, architecture, and output are unknown to the attacker.

The purpose of this thesis is to explore the applicability of certain classes of testing attacks on real-world deployed AI systems. For simplicity, many kinds of SOTA adversarial attacks have only been explored in the testing regime but have not been applied to truly black-box systems.

The main aim of the thesis will be to test different types of testing attacks on AI SaaS providers like Google Vision API, Amazon recognition, or Clarify. Understandably, attacking an unknown system will be more challenging than attacking a known neural network in the testing stage. We will try to measure this attack difficulty increase. This information could prove helpful in selecting the most promising attack to a specific situation.

Many SOTA testing attacks were not designed to attack specific deployed systems, so some attacks will need to be slightly changed to be used. We will explore different ways to modify existing attacks and evaluate them.

If some of those services are proven to be vulnerable, this would have a massive impact on all downstream applications using those SaaS APIs. For instance, content moderation mechanisms, which rely mainly on automatic detection, could be circumvented.

2. Preliminaries

In this section we introduce and explain the necessary theoretical background, upon which we build later on.

We first define the concepts of Machine learning (ML) and Deep neural networks (DNNs). We explain, how DNNs can learn patterns from data by using powerful gradient-based stochastic optimization algorithm called Stochastic gradient descent (SGD). Then we talk about a subset of DNNs that perform very well on image data called Convolutional neural networks (CNNs). Finally, we describe the systemic vulnerability of DNNs and we define and explain adversarial attacks and defenses, adversarial examples, adversarial robustness and different adversarial attack threat models.

2.1   Machine learning

2.2   Deep neural networks

2.3   DNN training

2.4   Convolutional neural networks

2.5   Adversarial attacks on DNNs

2.6   Threat models

Anděl [2007]

3. Related Work

3.1 Title of the first subchapter of the third chapter

Conclusion

Bibliography

J. Anděl. *Základy matematické statistiky.* Druhé opravené vydání. Matfyzpress, Praha, 2007. ISBN 80-7378-001-1.

List of Figures

List of Tables

List of Abbreviations

A. Attachments

A.1   First Attachment