FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

BACHELOR THESIS

Štěpán Procházka

Adversarial examples generation for deep neural networks

Department of Theoretical Computer Science and Mathematical Logic

Supervisor of the bachelor thesis:	Mgr. Roman Neruda, CSc.
Study programme:	Computer Science
Study branch:	General Computer Science

Prague 2018

I declare that I carried out this bachelor thesis independently, and only with the cited sources, literature and other professional sources.

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 .... signature of the author

TODO Dedication.

Title: Adversarial examples generation for deep neural networks

Author: Štěpán Procházka

Department: Department of Theoretical Computer Science and Mathematical Logic

Supervisor: Mgr. Roman Neruda, CSc., Department of Theoretical Computer Science and Mathematical Logic

Abstract: TODO Abstract.

Keywords: machine learning adversarial examples evolutionary algorithms deep learning

Contents

Introduction

The effort to automate various processes in our lives has always been one of the key concepts of scientific research. The automation of mechanical work has already been solved to great extent by advances in engineering. On the contrary, automated processing of information has been solved just partially. Well defined tasks i.e., tasks with fully defined behaviour can be solved and the challenge lies just in effectivity of the solution. On the other hand, models solving tasks based on raw real word data, human level input and output or some degree of fuzziness are yet to be found or, if they exist, suffer from several shortcomings. One of those shortcomings is vulnerability to adversarial examples i.e., artificially created inputs misinterpreted by those models. In this thesis, we will cover the task of generating adversarial examples for the models used for classification in computer vision.

TODO 800 chars artificial intelligence -¿ machine learning -¿ sota deep learning

TODO 800 computer vision and classification tasks

The effort to automate various tasks in our lives has always been one of the most important concepts of scientific research. The inventions of various machines, from the water wheel through the steam engine to the internal combustion engine along with the electric engine, enabled mankind to automate mechanical work. In the same manner, people have always been seeking for ways to automate reasoning.

The Syllogisms defined by Aristotle are often considered to be the first attempts to automate reasoning. Formation of propositional logic together with predicate logic laid foundations to formal logic. The evolution of formal logic and its parts, namely Boolean algebra originated theoretical computer science and, consequently, enabled the first digital computers to be built. Among other computer science fields, artificial intelligence was born. Various approaches towards automated reasoning emerged over the last few decades. While logical programming, ruled based systems and other ??formal?? paradigms proved successfull in solving multiple well defined tasks, they rendered useless when solving tasks with raw real world input or fuzzy data. SVM??. With soaring computational capabilities machine learning became state of the art.

1. Underlying Theory

In this chapter we will present theoretical background of various ??subjects (general word for methods, models, structures etc...)?? relevant to this ??writing??. We will cover various Artificial Intelligence paradigms with emphasis on Deep Learning for image classification and image generation, and Evolutionary Algorithms as a state space search method. Adversarial Example generation methods will be discussed as well. TODO Overview - AI, ML, Optimization, Search - context and structure

- Artificial Intelligence (take out tasks, AI as methods) : Tasks — Classification - Regression - Sampling unknown distribution - Planning - Translation : Methods - Graph Search - Random/Beam/Local Search - Genetic Algorithms — Evolutionary Algorithms

- SVM : Machine Learning - Neural Networks — Deep Learning - Recurrent Neural Networks - Adversarial Examples

An example citation: Anděl [2007]

1.1 Classification Tasks in Computer Vision

TODO datasets and their properties

1.2 Deep Learning in Image Classification

1.3 Evolutionary Algorithms as Generative Method

TODO

1.4 Adversarial Examples for Deep Learning Models

TODO topology - black/white box, iterative/single-shot, targeted/non-targeted

2. State of The Art Approaches

2.1 Gradient based methods

2.2 Variational Auto-Encoders

3. Our Solution

3.1 Pure Evolutionary Algorithms

3.2 Hybrid Methods

4. Experiments

Experiments will be carried out for those scenarios:

- FGSM (white-box, target model)
- FGSM (surrogate, different seed model)
- FGSM (surrogate, teacher-student model)
- FGSM (surrogate, specific binary teacher-student model)
- EA (black-box, pure EA)
- hybrid (surrogate, EA + pre-trained surrogate)
- hybrid (surrogate, EA + on-demand-trained surrogate)

Each combination of following options will be tested:

- targeted vs non-targeted
- single image vs multi-image vs generalization (unseen images)

Each experiment will be cross-validated using 14-fold cross-validation.

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. Evgena Framework User Documentation

B. Attachments

B.1 First Attachment