

Distributed Adversarial Attacks

Sander Prenen

Thesis voorgedragen tot het behalen
van de graad van Master of Science
in de ingenieurswetenschappen:
computerwetenschappen, hoofdoptie
Artificiële intelligentie

Promotoren:

Prof. dr. ir. W. Joosen
Dr. ir. D. Preuveneers

Assessoren:

Ir. W. Eetveel
W. Eetrest

Begeleiders:

V. Rimmer
I. Tsingenopoulos

© Copyright KU Leuven

Without written permission of the thesis supervisors and the author it is forbidden to reproduce or adapt in any form or by any means any part of this publication. Requests for obtaining the right to reproduce or utilize parts of this publication should be addressed to the Departement Computerwetenschappen, Celestijnenlaan 200A bus 2402, B-3001 Heverlee, +32-16-327700 or by email info@cs.kuleuven.be.

A written permission of the thesis supervisors is also required to use the methods, products, schematics and programmes described in this work for industrial or commercial use, and for submitting this publication in scientific contests.

Zonder voorafgaande schriftelijke toestemming van zowel de promotoren als de auteur is overnemen, kopiëren, gebruiken of realiseren van deze uitgave of gedeelten ervan verboden. Voor aanvragen tot of informatie i.v.m. het overnemen en/of gebruik en/of realisatie van gedeelten uit deze publicatie, wend u tot het Departement Computerwetenschappen, Celestijnenlaan 200A bus 2402, B-3001 Heverlee, +32-16-327700 of via e-mail info@cs.kuleuven.be.

Voorafgaande schriftelijke toestemming van de promotoren is eveneens vereist voor het aanwenden van de in deze masterproef beschreven (originele) methoden, producten, schakelingen en programma's voor industrieel of commercieel nut en voor de inzending van deze publicatie ter deelname aan wetenschappelijke prijzen of wedstrijden.

Preface

Sander Prenen

List of Abbreviations

AI Artificial Intelligence. 3

ANN Artificial Neural Network. 3

ReLU Rectified Linear Unit. 3, 4

Contents

Preface	i
Contents	iii
List of Figures	iv
List of Tables	v
1 Introduction	1
2 Background	3
2.1 Neural networks	3
2.2 Adversarial attacks	3
2.3 Particle swarm optimization	3
Bibliography	5

List of Figures

2.1	Linear separability	4
2.2	Activation functions	4

List of Tables

Chapter 1

Introduction

Chapter 2

Background

2.1 Neural networks

Ever since the invention of computer systems, it has always been a goal of scientists and engineers to create **Artificial Intelligence (AI)**. Current state of the art approaches are mimicking the human brain, more specifically the neurons inside the brain. Already in the fifties, Rosenblatt introduced his perceptron [1]. The perceptron is a single neuron able to learn linearly separable patterns. It does so by finding a hyperplane that separates the two classes. This hyperplane is called the decision surface or decision boundary. The concept of linear separability is explained in Figure 2.1 in two dimensions.

Unfortunately not all patterns are linearly separable. To overcome this problem, the neurons can be layered, creating an **Artificial Neural Network (ANN)** in the process. Layering neurons sequentially is essentially a linear combination of neurons. This in itself does not create non-linear decision surfaces. Non-linear activation functions are added for the **ANN** to be able to learn more complex decision boundaries. Some commonly used activation functions are **Rectified Linear Unit (ReLU)** [2], Heaviside step function and softmax (or sigmoid when used on scalars). In Figure 2.2 the plots of the activation functions can be found.

2.2 Adversarial attacks

2.3 Particle swarm optimization

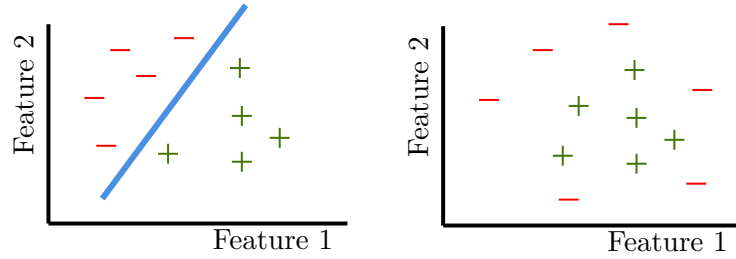


FIGURE 2.1: Linearly separable classes on the left and non-linearly separable classes on the right. Two classes are linearly separable if there exists a hyperplane for which all examples of one class are on the same side of this hyperplane, whilst all examples of the other class are on the other side of the hyperplane. In two dimensions, the hyperplane is a straight line.

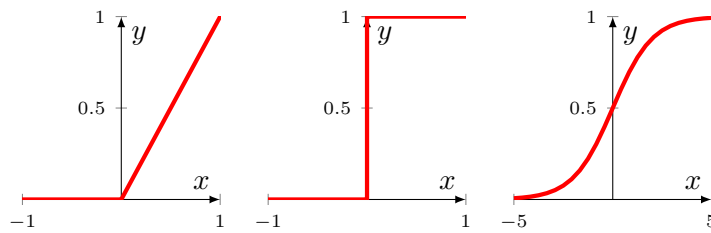


FIGURE 2.2: Plots of different activation functions. From left to right: **ReLU**, Heaviside step and sigmoid.

Bibliography

- [1] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. 65(6):386–408.
- [2] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML'10*, page 807–814, Madison, WI, USA, 2010. Omnipress.