SAPIENZA
UNIVERSITÀ DI ROMA

Robustness Of Deep Neural Networks Using Trainable Activation Functions

Computer Science - Informatica LM-18

Corso di Laurea Magistrale in Computer Science

Candidate

Federico Peconi

ID number 1823570

Thesis Advisor

Prof. Simone Scardapane

Academic Year 2019/2020

Thesis defended on Something October 2020
in front of a Board of Examiners composed by:

Prof. Nome Cognome (chairman)

Prof. Nome Cognome

Dr. Nome Cognome

Robustness Of Deep Neural Networks Using Trainable Activation Functions
Master's thesis. Sapienza – University of Rome

This thesis has been typeset by LaTeX and the Sapthesis class.

Version: August 30, 2020

Author's email: peconi.1823570@studenti.uniroma1.it

Dedicated to
Donald Knuth

Abstract

This document is an example which shows the main features of the LaTeX 2_ε class `sapthesis.cls` developed by with the help of GuIT (Gruppo Utilizzatori Italiani di TeX).

Acknowledgments

Ho deciso di scrivere i ringraziamenti in italiano per dimostrare la mia gratitudine verso i membri del GuIT, il Gruppo Utilizzatori Italiani di TEX, e, in particolare, verso il prof. Enrico Gregorio.

Contents

Chapter 1

Introduction

1.1 Intriguing Properties of Neural Networks

Here we informally state the problem of adversarial attacks in ML models especially wrt to Neural Networks. Why is it of fundamental importance for the progress of the field from both practical (nns cant yet be deployable in critical scenarios for such reasons) and theoretical (Madry arguments around interpertability and robustness) perspectives

1.2 Smooth Activation Functions and Robustness

Recently a link has been proposed between activation functions and the robustness of Neural Networks (Smooth Adversarial Training). In particular, authors showed how they managed to improve the robustness by replacing the traditional Rectified Linear Units activation functions with smoother alternatives such as ELUs, SWISH, PReLUs

Building up from this result we thought we could find benefits by laveraging recently proposed smooth trainable activation functions called Kernel Based Activation Functions (Scardapane et al.), which already showed great results in standard tasks, in the context of adversarial attacks.

1.3 Structure of the Thesis

Description of the remaining chapters

Chapter 2

Fundamentals

2.1 Deep Neural Networks

Broadly speaking, the field of Machine Learning is the summa of any algorithmic methodology whose aim is to automatically find meaningful patterns inside data without being explicitly programmed on how to do it. Well known examples are: Search Trees (ref.), Support Vector Machines (ref.), Clustering (ref.) and, more recently, Neural Networks (ref.). During the last two decades Neural Networks gained a lot of attention for their outstanding performances in different tasks like image classification (ref. ImageNet, over human level), speech and audio processing(ref).

2.1.1 Definition

Neural Networks are often used in the context of Supervised Learning where the objective is to model a function $\hat{f} : X \mapsto Y$

2.1.2 Training

2.1.3 CNNs: Convolutional Neural Networks

2.1.4 From Neural Networks to Deep Neural Networks

2.2 Adversarial Examples Theory

Formal definition of what is an adversarial attacks plus currently well known attacks

2.3 Defenses

Review of the literature on defenses to improve robustness: provable robustness, adversarial training

2.4 Kernel Based Activation Functions

Chapter 3

Related Works

3.1 K-Winners Take All

3.2 Smooth Adversarial Training

Chapter 4

Solution Approach

Comparing the activations's distributions for different activation functions (ReLU, KWTA, Kafs) seem to suggest Kafs might be good candidates to improve model robustness

4.1 Lipschitz Constant Approach

On the limitations of current Lipischitz-Constant based approaches especially when involving Kafs

4.2 Fast is Better than Free Adversarial Training

Adversarial training (Madry et al.) and current methods to improve the efficiency (Fast is better than free)

Chapter 5

Evaluation

5.1 VGG Inspired Architectures Results

5.2 Explofing Gradients with KafNets

The exploding gradients problem with KafResNet, why is it happening? (still to clarify)

5.3 ResNet20 Inspired Architectures Results

Chapter 6

Future Works

Different Kernels, resolve the exploding gradient problem and scale to ImageNet
Perform more adaptive attacks to assess the robustness of kafresnets as is the current standard (Carlini et al.)

Chapter 7

Conclusions

This thesis tries to add to the bag of evidences in literature that smoother architectures might benefit improvements in adversarial resiliency