Compression to the Rescue: Defending from Adversarial Attacks Across Modalities

Extended Abstract

Nilaksh Das  
Georgia Institute of Technology  
School of CSE  
Atlanta, GA, USA  
nilikshdas@gatech.edu

Madhuri Shanbhogue  
Georgia Institute of Technology  
College of Computing  
Atlanta, GA, USA  
madhuri.shanbhogue@gatech.edu

Shang-Tse Chen  
Georgia Institute of Technology  
College of Computing  
Atlanta, GA, USA  
schen351@gatech.edu

Fred Hohman  
Georgia Institute of Technology  
School of CSE  
Atlanta, GA, USA  
fredhohman@gatech.edu

Siwei Li  
Georgia Institute of Technology  
College of Computing  
Atlanta, GA, USA  
robertsiweili@gatech.edu

Li Chen  
Intel Corporation  
Intel Labs  
Hillsboro, OR, USA  
li.chen@intel.com

Michael E. Kounavis  
Intel Corporation  
Intel Labs  
Hillsboro, OR, USA  
michael.e.kounavis@intel.com

Duen Horng Chau  
Georgia Institute of Technology  
School of CSE  
Atlanta, GA, USA  
polo@gatech.edu

ABSTRACT

Research in the upcoming field of adversarial ML has revealed that machine learning, especially deep learning, is highly vulnerable to imperceptible adversarial perturbations, both in the domain of vision as well as speech. This has induced an urgent need to devise fast and practical approaches to secure deep learning models from adversarial attacks, so that they can be safely deployed in real-world applications. In this showcase, we put forth the idea of compression as a viable solution to defend against adversarial attacks across modalities. Since most of these attacks depend on the gradient of the model to craft an adversarial instance, compression, which is usually non-differentiable, denies a useful gradient to the attacker. In the vision domain we have JPEG compression, and in the audio domain we have MP3 compression and AMR encoding — all widely adopted techniques that have very fast implementations on most platforms, and can be feasibly leveraged as defenses. We will show the effectiveness of these techniques against adversarial attacks through live demonstrations, both for vision as well as speech. These demonstrations would include real-time computation of adversarial perturbations for images and audio, as well as interactive application of compression for defense. We would invite and encourage the audience to experiment with their own images and audio samples during the demonstrations. This work was undertaken jointly by researchers from Georgia Institute of Technology and Intel Corporation.

CCS CONCEPTS

• Computing methodologies → Speech recognition; Computer vision; Image compression;

KEYWORDS

adversarial ML, computer vision, speech recognition, deep learning

OVERVIEW

The threat of adversarial attack casts a shadow over deploying Deep Neural Networks (DNNs) in security and safety-critical applications, such as autonomous vehicles. In computer vision applications such as image classification, an attacker can add visually imperceptible perturbations to each pixel of an image and mislead a DNN model into making arbitrary predictions [4]. In automatic speech recognition (ASR) systems, the attacker can manipulate an audio sample by carefully introducing faint "noise" in the background that humans easily dismiss. Such perturbation causes the ASR model to transcribe the manipulated audio sample as a target phrase of the attacker’s choosing [1]. Given these vulnerabilities, there is an urgent need to resolve these threats with fast and practical approaches to secure vision and speech models from such attacks. In this showcase, we present SHIELD [3] and Adagio [2] — two frameworks for the defense of vision and speech models respectively. These frameworks were developed at Georgia Institute of Technology in collaboration with researchers from Intel Labs, specifically keeping practicality of real-life deployment in mind. In both of these works, we leverage the idea that compression — a central concept that underpins numerous successful data mining techniques — can offer powerful, scalable, and practical defense for deep learning in real-time.

SHIELD

SHIELD stands for Secure Heterogeneous Image Ensemble with Localized Denoising. It is a multi-faceted framework that combines randomization, compression, re-training and ensembling into
Figure 1: Screenshot of **ADAGIO** with an example usage scenario. (1) Jane uploads an audio file that is transcribed by DeepSpeech (an ASR model); then she performs an adversarial attack on the audio in real time by entering a target transcription after selecting the attack option from the dropdown menu. (2) Jane decides to perturb the audio to change the last word of the sentence from “joanna” to “marissa”; she can listen to the original audio and see the transcription by clicking on the “Original” badge. (3) Jane applies MP3 compression to recover the original, correct transcription from the manipulated audio; clicking on a waveform plays back the audio from the selected position. (4) Jane can experiment with multiple audio samples by adding more cards.

ADAGIO

**ADAGIO (Adversarial Defense for Audio in a Gadget with Interactive Operations)** is another interactive tool that allows real-time computation of adversarial attack on a speech-to-text model. It incorporates MP3 compression and AMR encoding as defenses. In [2], we show that these techniques are able to effectively eliminate targeted attacks, reducing the attack success rate from 92.5% to 0%. An example of the interactive demo which we plan to show the audience for this system can be viewed at https://youtu.be/0W2BKMwSfVQ.

**REFERENCES**


