Radboud University Nijmegen

Faculty of Science

Backdoor attack on deep neural networks using inaudible triggers

Dolphin attack trigger

Thesis BSc Computing Science

Author:
Julian van der Horst

Supervisor:
Stjepan Picek
Stefanos Koffas

December 2022

Contents

1 Introduction **2**

2 Background **2**
2.1 Automatic Speech Recognition (ASR) 2
2.2 Backdoor attacks 2
2.3 Microphone 2
2.4 BackDoor 2
2.5 Threat model 2

3 Experimental setup **2**
3.1 The data and parameters 2
3.2 The convolutional neural network 2
3.2.1 The poison data 2
3.3 The signal-producing device 2
3.4 The speech recognition app 2
3.4.1 The phone 2

4 Experiment **2**

5 Conclusion **2**

1 Introduction

2 Background

2.1 Automatic Speech Recognition (ASR)

Automatic Speech Recognition, otherwise known as ASR, has been around since 1952 when bell labs were able to recognize digits spoken over the phone [4]. Back then, analog circuitry was used to understand the incoming signal and identify a digit. Nowadays, these analog circuits are replaced by deep learning models where. They take audio in a compressed form to train and then recognize speech. One of these compressed forms is MFCC (Mel-frequency cepstral coefficient), which was invented in the 1980s and is still widely used today. I used MFCCs in my convolutional neural network in my research since it focuses on information from human speech and deemphasizes other information [1].

2.2 Backdoor attacks

Introduce the idea of a backdoor attack and especially with audio neural networks

2.3 Microphone

Explain shortly how modern microphones work and why a MEMS microphone is special

2.4 BackDoor

[?]

Explain the idea of the BackDoor paper and how we will create the trigger

2.5 Threat model

Explain the transmitter, receiver and gray box data poisoning. Also, add

3 Experimental setup

3.1 The data and parameters

Chapter 7.1 [2] [5] [3]

Show that we used the standard tensorflow speech commands dataset with 9 commands. Also show that we used a

3.2 The convolutional neural network

3.2.1 The poison data

3.3 The signal-producing device

3.4 The speech recognition app

3.4.1 The phone

4 Experiment

5 Conclusion

References

[1] Namrata Dave. Feature extraction methods lpc, plp and mfcc in speech recognition. *International journal for advance research in engineering and technology*, 1(6):1–4, 2013.

[2] Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon. Spoken language processing: A guide to theory, algorithm, and system development. 01 2001.

[3] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and X. Zhang. Trojaning attack on neural networks. In *Network and Distributed System Security Symposium*, 2018.

[4] Douglas O'Shaughnessy. Invited paper: Automatic speech recognition: History, methods and challenges. *Pattern Recognition*, 41(10):2965–2979, 2008.

[5] Saeid Samizade, Zheng-Hua Tan, Chao Shen, and Xiaohong Guan. Adversarial example detection by classification for deep speech recognition. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3102–3106, 2020.