POSTER: Detecting Audio Adversarial Example through Audio Modification

Hyun Kwon
Korea Advanced Institute of Science and Technology
School of Computing
Daejeon, South Korea
khkh@kaist.ac.kr

Hyunsoo Yoon
Korea Advanced Institute of Science and Technology
School of Computing
Daejeon, South Korea
hyoon@kaist.ac.kr

Ki-Woong Park*
Sejong University
Department of Computer and Information Security
Seoul, South Korea
woongbak@sejong.ac.kr
*Corresponding author

ABSTRACT

Deep neural networks (DNNs) perform well in the fields of image recognition, speech recognition, pattern analysis, and intrusion detection. However, DNNs are vulnerable to adversarial examples that add a small amount of noise to the original samples. These adversarial examples have mainly been studied in the field of images, but their effect on the audio field is currently of great interest. For example, adding small distortion that is difficult to identify by humans to the original sample can create audio adversarial examples that allow humans to hear without errors, but only to misunderstand the machine. Therefore, a defense method against audio adversarial examples is needed because it is a threat in this audio field. In this paper, we propose a method to detect audio adversarial examples. The key point of this method is to add a new low level distortion using audio modification, so that the classification result of the adversarial example changes sensitively. On the other hand, the original sample has little change in the classification result for low level distortion. Using this feature, we propose a method to detect audio adversarial examples. To verify the proposed method, we used the Mozilla Common Voice dataset and the DeepSpeech model as the target model. Based on the experimental results, it was found that the accuracy of the adversarial example decreased to 6.21% at approximately 12 dB. It can detect the audio adversarial example compared to the initial audio sample.

KEYWORDS

Deep neural network (DNN), Audio adversarial example, Defense method, Audio modification

ACM Reference format:
Hyun Kwon, Hyunsoo Yoon, and Ki-Woong Park*. 2019. POSTER: Detecting Audio Adversarial Example through Audio Modification. In *Proceedings of 2019 ACM SIGSAC Conference on Computer and Communications Security, London, United Kingdom, November 11–15, 2019 (CCS '19),* 3 pages.
https://doi.org/10.1145/3319535.3363246

CCS '19, November 11–15, 2019, London, United Kingdom

ACM ISBN 978-1-4503-6747-9/19/11.
https://doi.org/10.1145/3319535.3363246

1 INTRODUCTION

Deep neural networks (DNNs) [1] are used for machine learning tasks such as image recognition, speech recognition, pattern analysis, and intrusion detection. However, DNNs are vulnerable to adversarial examples [2] that add a little amount of noise to the original sample. For example, if an attacker adds some noise to a turn left road sign, the modified turn left road sign is correctly recognized by a person but incorrectly recognized as a turn right road sign by a DNN-equipped autonomous vehicle. This type of an adversarial example leads to a decrease in the DNN's performance, and therefore, significant amount of research is being conducted in the field of image processing to address the aforementioned issue.

Adversarial examples have also recently been extended to the domain of audio, and several papers [3] [4] [5] [6] [7] have been introduced presenting different approaches to combat the threat. Vaidya et al. [3] proposed using the cocaine noodles method that will mislead a speech recognition system by making a strange sound, which cannot be understood by a person. To improve cocaine noodles method, Carlini et al. [4] suggested a hidden voice command method to improve the strange sound that people can not understand by adding human feedback. Zhang et al. [5] proposed the dolphin attack method, which makes the speech recognition system misleading by producing a high frequency band that cannot be heard by humans. Carlini and Wagner (CW) [6] recently published a paper on an attack that generates an audio adversarial example by adding a little BIT of noise to the original sample. This method improves the CTC loss function [8] by adding a little BIT of noise to the original voice data such that it is not mistaken by a human but is mistaken by the recognition system. As described above, these are some of attacks on adversarial examples in the audio field, but research on defense methods is also needed.

We propose a detection method that can reduce the effect of a CW attack, a state-of-the-art attack on the DeepSpeech model [9]. Our method applies the difference between the classification results of the original image and the adversarial example through audio modification. The contributions of this paper in the field of adversarial examples through audio modification are as follows:

- This is the first study that focuses on detecting audio adversarial examples using audio modification. We systematically show the platform of the proposed method.
- We analyze the spectrum, waveform, and accuracy rate of the proposed method. We also present the possibility of various ensemble methods of audio modification.

- We show the performance of the proposed method for the state-of-the-art DeepSpeech and the state-of-the-art CW attack.

The composition of this paper is as follows. In Section 2, the proposed method is presented. The evaluation is described in Section 3. Section 4 concludes the paper.

2 METHODOLOGY

The CW attack methodology used in the scenario is described in Section 2.1. Section 2.2 introduces the proposed defensive method.

2.1 Attack method

The CW attack [6] is a state-of-the-art attack that generates adversarial examples with 100% attack success rate.

$$\text{minimize } dB_x(x, x^*) + \sum_i c_i \times g_i(x^*, \pi^i), \tag{1}$$

where $dB_x(x, x^*)$ is a distortion loss function between adversarial example x^* and original sample x. $\sum_i c_i \times g_i(x^*, \pi^i)$ is a loss function of the sequence. DeepSpeech misclassifies x^* as target phrase t due to the loss function of the sequence. $dB_x(x, x^*)$ is the distortion loss function between adversarial example x^* and original sample x. CW attack generates an audio adversarial example that is misclassified as the target phrase by DeepSpeech, while minimizing the distortion and adjusting the c value.

2.2 Proposed method

Figure 1 shows an overview of the proposed method comprising two steps. First, the initial audio sample is verified against the recognition system and given an initial classification result. Subsequently, a modified audio signal is generated by audio modification of the initial audio sample. Next, the generated modified audio signal is compared with the classification result of the initial audio sample. If the difference in the classification result is large, the initial audio sample is regarded as an adversarial example. If the difference is small, the initial audio sample is regarded as an original sample.

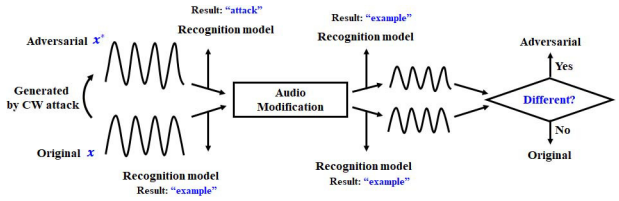


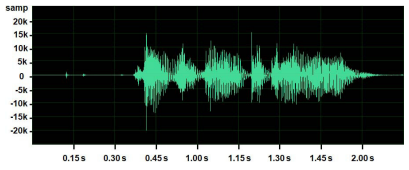
Figure 1: Overview of the proposed method

In principle, our method employs the features of the audio adversarial example. In the process of creating an audio adversarial example, some distortion is added to the original sample to a point where the machine begins to misinterpret the signal. Therefore, if the distortion is caused by audio modification, the difference in the classification result by the adversarial example is larger than that of the original sample.

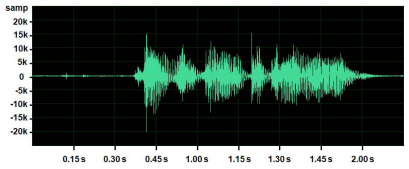
For audio modification, low-pass filters [10], high-pass filters [11], or notch filters [12] can be used, or any combination of the aforementioned filters can be used with equal success. Thus, a variety of audio modification combinations can lead to an improvement in the detection of audio adversarial examples through comparative analysis. In this paper, we applied the low-pass filter method [10] as a single method.

3 EVALUATION

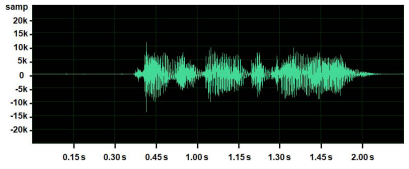
We used 100 arbitrary samples of the Mozilla Common Voice dataset [13] during our evaluation. The average time for each sample was about 6 s and the dataset was 16bit with 16000 Hz. We used a Tensorflow [14] library and an Intel(R) i5-7100 3.90-GHz server. We used pretrained DeepSpeech models [9] with an 83.51% accuracy rate for the audio recognition system and a low-pass filter method. The CW attack method was used to generate the audio adversarial examples. The learning rate was at 10, and Adam [15] was used as the optimizer.



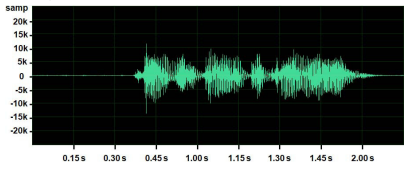
(a) Original sample (before audio modification)



(b) Adversarial example (before audio modification)



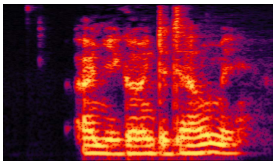
(c) Original sample (after audio modification)



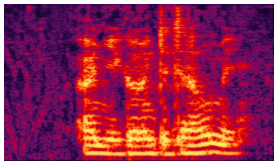
(d) Adversarial example (after audio modification)

Figure 2: An original sample and an audio adversarial example (before and after audio modification).

In terms of experimental results, Figure 2 shows the waveform before and after audio modification for the original sample and audio adversarial example. Figure 2 (b) shows a little BIT noise added overall in Figure 2 (a). In particular, when looking at the spectrum in Figure 3, Figure 2 (b) shows overall noise compared to Figure 2 (a). Therefore, the audios of Figures 2 (a) and 2 (b)



(a) Original sample (before audio modification)



(b) Adversarial example (before audio modification)

Figure 3: Spectrum of cases (a) and (b) of Figure 2.

Original sentence: "aren't you going to tell me"
Transcription of Figure 2 (a) and (c): "an y going to tell me"
Transcription of Figure 2 (b): "example"
Transcription of Figure 2 (d): "nd e going to tell me"

Figure 4: Transcription of DeepSpeech in Figure 2.

are almost the same. On the other hand, Figure 4 shows that the recognition system correctly recognizes an audio of Figure 2 (a) as "an y going to tell me", but misrecognizes an audio of Figure 2 (b) as "example" chosen by the attacker. However, the samples modified by audio modification are similar to the original sentence as shown in Figures 2 (c) and 2 (d). In view of the original sample in Figures 2 and 4, Figures 2 (a) and 2 (c) show some waveforms that are different but have the same interpretation. However, from the perspective of the adversarial example in Figures 2 and 4, it can be seen that Figure 2 (d) changes similar to the original sentence due to audio modification effects by removing the adversarial noise.

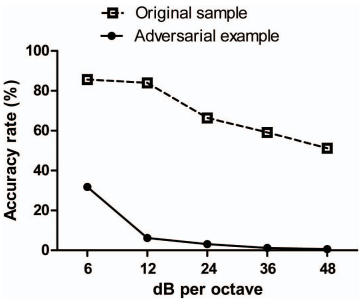


Figure 5: Accuracy rate of original sample and adversarial example through audio modification (roll-off: dB per actave).

Figure 5 shows the accuracy rate of the original sample and the adversarial example over the dB range of an octave in the low-level method. We tested 100 samples and compared how the accuracy of the concordance rate of the original sentence in the original sample varied with the concordance rate of the attack sentence in the adversarial example. Figure 5 demonstrates that the accuracy rate of the adversarial example decreases while maintaining the accuracy of the original sample until value of dB is 12. However, if the number of dB increases to be over 24, the accuracy of the original sample is reduced due to severe audio modification. Therefore, at exactly 12 dB, it can be considered to have hit a "sweet spot".

4 CONCLUSION

In this paper, we proposed a method to detect audio adversarial examples through audio modification. The key point of this method was the addition of a new low level distortion using audio modification, enabling classification result of the adversarial example to changes in sensitivity. In contrast, the original sample undergoes only a slight change in the classification result for low level distortion. Experimental results show that the accuracy of the adversarial example decreases to 6.21% at approximately 12 dB. It can detect the audio adversarial example compared to the initial audio sample. Future research will focus on the examination of our defensive method in terms of an ensemble strategy.

ACKNOWLEDGMENT

This work was supported by Institute for Information & communications Technology Promotion (No.2018-0-00420) and supported by the National Research Foundation of Korea (2017R1C1B2003957, 2017R1A2B4006026).

REFERENCES

[1] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.
[2] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. URL http://arxiv.org/abs/1312.6199.
[3] Tavish Vaidya, Yuankai Zhang, Micah Sherr, and Clay Shields. Cocaine noodles: exploiting the gap between human and machine speech recognition. *WOOT*, 15: 10–11, 2015.
[4] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. Hidden voice commands. In *USENIX Security Symposium*, pages 513–530, 2016.
[5] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. Dolphinattack: Inaudible voice commands. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 103–117. ACM, 2017.
[6] Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. *Deep Learning and Security Workshop*, 2018.
[7] Hyun Kwon, Yongchul Kim, Hyunsoo Yoon, and Daeseon Choi. Selective audio adversarial example in evasion attack on speech recognition system. *IEEE Transactions on Information Forensics and Security*, 2019.
[8] Beth Logan et al. Mel frequency cepstral coefficients for music modeling. In *ISMIR*, volume 270, pages 1–11, 2000.
[9] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.
[10] Fuchao Wang, Dapeng Tian, and Yutang Wang. High accuracy inertial stabilization via kalman filter based disturbance observer. In *2016 IEEE International Conference on Mechatronics and Automation*, pages 794–802. IEEE, 2016.
[11] Thoriq Bayu Aji, Jangkung Raharjo, and Ledya Novamizanti. Analisis audio watermarking berbasis dwt dengan metode qr decomposition dan quantization index menggunakan pso: Analysis audio watermarking based on dwt using qr decomposition and quatization index use pso. *eProceedings of Engineering*, 6(1), 2019.
[12] Robert C Maher. Audio signal enhancement. In *Principles of Forensic Audio Analysis*, pages 69–84. Springer, 2018.
[13] https://voice.mozilla.org/.
[14] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.
[15] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *The International Conference on Learning Representations (ICLR)*, 2015.