Attacking Audio Event Detection Deep Learning Classifiers with White Noise

Rodrigo dos Santos
rodrigoaugusto.silvadossantos@uta.edu
The University of Texas at Arlington
USA

Ashwitha Kassetty
ashwitha.kassetty@mavs.uta.edu
The University of Texas at Arlington
USA

Shirin Nilizadeh
shirin.nilizadeh@uta.edu
The University of Texas at Arlington
USA

ABSTRACT

We develop deep learning-based classifiers for Audio Event Detection (AED), attacking them next with some white noise disturbances. We show that an attacker can use such simple disturbances to potentially fully avoid detection by AED systems. Prior work has shown that attackers can mislead image classification tasks, however this work focuses on attacks against AED systems, by tampering the audio and not image. This work brings awareness to the designers and manufacturers of AED systems and devices, as these solutions are becoming more ubiquitous by the day.

CCS CONCEPTS

• **Security and privacy** → *Domain-specific security and privacy architectures.*

KEYWORDS

AED, neural networks, deep learning, spectrograms

ACM Reference Format:
Rodrigo dos Santos, Ashwitha Kassetty, and Shirin Nilizadeh. 2021. Attacking Audio Event Detection Deep Learning Classifiers with White Noise. In *The 14th PErvasive Technologies Related to Assistive Environments Conference (PETRA 2021), June 29-July 2, 2021, Corfu, Greece.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3453892.3464893

1 INTRODUCTION

Safety is a major concern in people's lives. Gun shooting represents one of the major threats to safety every person is exposed to. Situations like the Las Vegas Mandalay Bay hotel massacre, where a shooter fired his guns at defenseless and innocent country music concertgoers, killing 58 and harming over 850 people, are good examples of how everything can suddenly run out of control, bringing major impact to the lives of individuals, families and authorities.

Audio Event Detection Systems have the capability of capturing audio from the environment and leveraging some algorithm for detecting the presence of a specific sound of interest. AED systems have been employed for safety purposes, through the detection of suspicious sounds such as gunshots, footsteps and others. AED systems for detecting gunshot sounds make extensive use of state-of-the-art deep learning classifiers, such as Convolutional Neural Networks (CNN) [9] and Convolutional Recurrent Neural Networks (CRNN) [6], as their primary detection and classification algorithms.

A direct consequence of deep learning popularization was a proliferation of studies focusing on how to attack deep learning classifiers and systems enabled by them. We focus on this niche, and as such, study how to attack deep learning AED systems, focusing on employing simple, accessible and easy to reproduce disturbances made of white noise, to be used as a means to disrupt the classifiers. We focus on attacking the audio portion of the system, prior to image portion based on spectrogram generation, and to the best of our knowledge, no other work has focused on AED applications against audio disturbances under such research design constraints.

AED systems for gunshot detection can be employed anywhere from home to business and even public spaces, where they would constantly monitor the environment for suspicious events. In our threat model, we assume that the attacker, while attempting to cause harm, actively adds white noise perturbations to the environmental sound being captured by the AED system, overlaying it to the gunshot sounds being fired. The ultimate goal of the attacker is to prevent the AED system from detecting the gunshot sounds.

We implemented a CNN and a CRNN, and we use gunshot sounds datasets from [1, 2, 8]. We first tested the classifiers with undisturbed gunshot samples in order to examine their performance under baseline conditions, and then digitally injected white noise perturbations, interleaving them with the gunshot sounds, thus simulating an in the open scenario where these classifiers would be deployed as part of a suspicious sound detection solution.

Our consolidated results show that AED classifiers are susceptible against adversarial examples, as the performance of both the CNN and CRNN were strongly affected, being degraded by nearly 100% when tested against the perturbations. In particular, our contributions reside on attacking deep learning classifiers with a simple and very easy to reproduce disturbance, which is relevant to present day when there is a proliferation of real deep learning devices that rely on deep learning classifiers for suspicious sound detection.

2 METHODOLOGY

2.1 Neural Networks

Our neural networks take images of audio, or spectrograms, as their inputs. These display in a graph (usually 2D) the spectrum of frequency changes over time for a sound signal, by chopping it up and then stacking the slices one close to each other [7]. Our CNN presents three convolutional blocks with convolutional 2D layers and a total of 480 filters of size 3 by 3. It also presents two dense layers, employing ReLU and Softmax activations, besides sparse categorical cross entropy as loss function and RMSprop as optimizer. Our CRNN is composed by one convolutional block, with

PETRA 2021, June 29-July 2, 2021, Corfu, Greece
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8792-7/21/06.
https://doi.org/10.1145/3453892.3464893

one convolutional layer. This block is made by 128 filters of size 32, ReLU activation and batch normalization, one backwards LSTM layer with 128 units, followed by two stacked dense layers. We use a combination of tahn, ReLU and Softmax, and also sparse categorical cross entropy as loss function and Adam as optimizer.

2.2 Attacks with Noise

To attack the classifiers we employ *white noise*, which happens when each audible frequency is equally loud, so no sound feature, shape or form can be distinguished [3]. We selected this noise variant given its ubiquity in day-to-day life, and specially its simplicity with regards on how to reproduce it. It is important to highlight that attack occurs during audio capturing, prior to spectrogram generation, and as such, is an audio based attack.

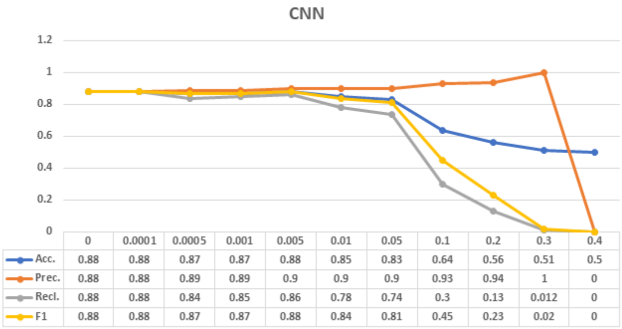
2.3 Experiments

Our experiments involve the use of our two neural network classifiers set as two different representation of an audio monitoring home security system. We employ digital gunshot samples, first in unnoisy conditions, and then we infuse the same samples with progressively higher levels noise.

1. **Unnoisy Experiments**: Both AED classifiers exposed to digital gunshot sounds, without any disturbance.
2. **White Noise Experiments**: Both AED classifiers exposed to digital gunshot sounds, now interleaved with increasing white noise levels, ranging from 0.0001 to 0.5.

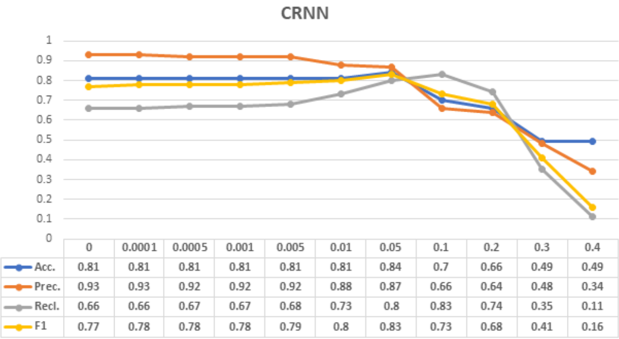
The unnoisy experiments generate our baselines, and both models perform reasonably. When we proceed to attack these classifiers with white noise, both models present drops in classification performance as soon as such noise is introduced to the test sets. The drops are small but cumulative, and a sharper drop is noticed when the 0.1 threshold is reached, only to get unacceptably worse from there on, to the point of rendering both models pretty much useless.

Easy to realize is that the CRNN proved to be slightly more robust than the CNN, and we credit this to its memory advantage over the CNN [4, 5]



	0	0.0001	0.0005	0.001	0.005	0.01	0.05	0.1	0.2	0.3	0.4
Acc.	0.88	0.88	0.87	0.87	0.88	0.85	0.83	0.64	0.56	0.51	0.5
Prec.	0.88	0.88	0.89	0.89	0.9	0.9	0.9	0.93	0.94	1	0
Recl.	0.88	0.88	0.84	0.85	0.86	0.78	0.74	0.3	0.13	0.012	0
F1	0.88	0.88	0.87	0.87	0.88	0.84	0.81	0.45	0.23	0.02	0

Figure 1: CNN classification performance on noise-free test datasets followed by increasing levels of white noise



	0	0.0001	0.0005	0.001	0.005	0.01	0.05	0.1	0.2	0.3	0.4
Acc.	0.81	0.81	0.81	0.81	0.81	0.81	0.84	0.7	0.66	0.49	0.49
Prec.	0.93	0.93	0.92	0.92	0.92	0.88	0.87	0.66	0.64	0.48	0.34
Recl.	0.66	0.66	0.67	0.67	0.68	0.73	0.8	0.83	0.74	0.35	0.11
F1	0.77	0.78	0.78	0.78	0.79	0.8	0.83	0.73	0.68	0.41	0.16

Figure 2: CRNN classification performance on noise-free test datasets followed by increasing levels of white noise

3 CONCLUSIONS

We tested CNN and CRNN algorithms for AED, and while their detection performance was reasonable under ideal circumstances, a sharp drop in it was seen, even when little white noise was injected into the test audio samples. This is important because white noise is simple to reproduce and to be employed by non-technically savvy individuals. It also can be hard to be filtered out without affecting the sound capture capability needed for an AED system, especially when higher noisy thresholds are used and when its amplitude is tailored to closely follow that of the sound of interest.

We do not believe to be far-fetched the envisioning of a scenario where malicious individuals plan in advance to carryout a gun-based attack, and in order to prevent or affect possible gunshot detection systems, also to make use of some medium to large scale white-noise reproducing gear based on large speakers and other specialized equipment. As such, our motivation is to be one step ahead of potential exploits. We also seek to motivate fellow researchers from the academy and professionals from the industry to think of potential security shortfalls before executing the design and the implementation of AED solutions.

REFERENCES

[1] AirborneSound. 2017. The Free Firearm Library - Expanded Edition. https://www.airbornesound.com.
[2] DCASE17. 2017. Detection of rare sound events. http://www.cs.tut.fi/sgn/arg/dcase2017.
[3] Ernest Edmonds. 2006. Abstraction and interaction: an art system for white noise. In *International Conference on Computer Graphics, Imaging and Visualisation (CGIV'06)*. IEEE, 423–427.
[4] Xinyu Fu, Eugene Ch'ng, Uwe Aickelin, and Simon See. 2017. CRNN: a joint neural network for redundancy detection. In *2017 IEEE International Conference on Smart Computing (SMARTCOMP)*. IEEE, 1–8.
[5] Shao En Gao, Bo Sheng Lin, and Chuin-Mu Wang. 2018. Share price trend prediction using CRNN with LSTM structure. In *2018 International Symposium on Computer, Consumer and Control (IS3C)*. IEEE, 10–13.
[6] Hyungui Lim, Jeongsoo Park, and Y Han. 2017. Rare sound event detection using 1D convolutional recurrent neural networks. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop*. 80–84.
[7] Roelandts. 2013. What is a spectrogram. https://tomroelandts.com/articles/what-is-a-spectrogram.
[8] Justin Salamon and Christopher Jacoby. 2014. A Dataset and Taxonomy for Urban Sound Research. http://www.justinsalamon.com.
[9] H. Zhou, Y. Song, and H. Shu. 2017. Using deep convolutional neural network to classify urban sounds. In *TENCON 2017 - 2017 IEEE Region 10 Conference*. 3089–3092.