Logit Noising Artifacts

1 Description

Logit Noising is audio adversarial examples detection system. This system is implemented based on DeepSpeech v0.1.1.

The artifact consists of one python file (**Detect_DeepSpeech.py**) and one jupyter notebook file (**logit_analysis_ACSAC.ipynb**).

Detect_DeepSpeech.py
This corresponds to Figure 4: Logit Noising Architecture. It is used to detect whether the input audio is benign or an adversarial example.

logit_analysis_ACSAC.ipynb
This corresponds to Section 4.1 (Difference in Logit Value Gap Distribution) and Section 4.2. (**Noise selection**). It is used to analyze logit difference between benign audio and audio adversarial example and select noise variable.

2 Installation

Download the **DeepSpeech v0.1.1** :
https://github.com/mozilla/DeepSpeech/tree/v0.1.1

Download the **pre-trained model(v0.1.0)** :
https://github.com/mozilla/DeepSpeech/releases/download/v0.1.0/deepspeech-0.1.0-models.tar.gz

Also, DeepSpeech requires native_client to execute ASR system. However, the native client for v0.1.1 is expired, so you should rebuild it by yourself.
Install **native_client**:
https://github.com/mozilla/DeepSpeech/tree/v0.1.1/native_client#readme

Download **Logit Noising code**:
https://github.com/namgyupark22/_Detecting_Audio_Adversarial_Examples_with_Logit_Noising

3 Requirement for reproducing Logit Noising

Detect_DeepSpeech.py

- Python 3.6, CUDA 8.0, and CUDNN 6.0
- one TITAN V GPU
- pip3 install editdistance (*as same as DeepSpeech v0.1.1)

logit_analysis_ACSAC.ipynb

- Python 3.6, CUDA 9.0, and CUDNN 7.6
- one TITAN V GPU
- pip3 install numpy scipy tensorflow-gpu==1.8.0 pandas python_speech_features, matplotlib

4 Reproduce Experiment

1. Install Logit noising code to DeepSpeech parent folder.
2. Move **Detect_DeepSpeech.py** to DeepSpeech folder.

3. Navigate DeepSpeech folder.

4. Run python3 -u Detect_DeepSpeech.py –checkpoint_dir **$pre-trained model**
- This will produce benign transcript result or detect audio adversarial examples

5. Make 3 types of attacks and evaluate 5 types attacks

Benign input data
- LibriSpeech :
https://www.openslr.org/resources/12/test-clean.tar.gz

Audio adversarial Examples
- Carlini and Wagner:
https://github.com/carlini/audio_adversarial_examples/tree/a8d5f675ac8659072732d3de2152411f07c7aa3a
- Hiromu:
https://github.com/hiromu/robust_audio_ae
- Taori:
https://github.com/rtaori/Black-Box-Audio
- Metamorph:
https://acoustic-metamorph-system.github.io
- Weighted-sampling:
https://sites.google.com/view/audio-adversarial-examples/