CS753 : Project Report

Adversarial Attack on ASR systems

Rohan Shah: 180070046
Jay Sawant: 18D070050

May 14, 2021

1 Introduction

Prior work [8] has shown that neural networks are vulnerable to adversarial examples , instances x' similar to a natural instance x, but classified incorrectly by a neural network. Until recently, a vast majority of the work was focused on the image domain, in areas such as image classification, segmentation, generative models etc.

Carlini and Wagner[1] demonstrated the existence of targeted adversarial examples in the audio domain. With the growing popularity of voice assistants such as Siri, Google now etc, and also the use of speaker verification as a form of identification, such attacks could have far reaching implications.

Our objective was to independently implement the algorithm proposed for English and Hindi targets. Given a natural waveform x and a target phrase t we attempt to construct an almost imperceptible perception δ such that $x + \delta$ is classified as t. We implemented the algorithm in PyTorch on a model trained on the Speech Commands Dataset. We successfully generated perturbations for targets both present and absent in the dataset.

2 Related Work

The adversarial examples produced by [1] can easily be distinguished from natural sound, and lose their adversarial property when played over air. In recent years there has been a surge of work in the field of audio adversarial attacks. [2] demonstrated the existence of imperceptible and robust attacks, which can be played over the air. [3] generated universal perturbations which can be applied to any input for a given model. In [4] the authors have shown that audio adversarial attacks are transferable, at least to some extent, demonstrating. reasonable success attacking Wave2Vec with adversarial samples generated on DeepSpeech.

While all the above attacks are white-box, i.e. they assume full access to the model and its gradients, the more realistic black-box setting has also received considerable interest. [7] demonstrated black-box attacks using genetic algorithms on state of the art Speech models.

With the rapid development of adversarial attacks, there has been a lot of work on building defenses as well. In the image domain, most defenses are based on input transformations. However, many existing methods are shown to be bypassed by subsequent or adaptive adversarial attacks (Carlini & Wagner, 2017b; He et al., 2017; Carlini & Wagner, 2017c; Lu et al., 2018). Similarly [6] finds that audio input transformations based on waveform quantization, temporal filtering, signal downsampling or autoencoder reformation suffers from similar weakness. [6] also finds that temporal dependency properties can be used effectively to defend against audio attacks, even when the adversary has knowledge of the defense applied. [5] proposes a randomized defense method which uses psychoacoustic models similar to those used in the attacks.

3 Methodology

Threat Model: We assume a white-box setting where we have full access to the model, and can backpropogate gradients through it. However the model parameters remain frozen throughout the process. We accept only those adversarial examples which exactly match the target phrase.

Distortion: We seek to minimize the distortion produced, δ while generating our adversarial examples. We define $dB_x(\delta) = dB(\delta) - dB(x)$. We attempt to find the adversarial example with a minimum value of $dB_x(\delta)$.

Formulation:

minimize $|\delta|_2^2 + c.l(x+\delta, t)$
such that $dB_x(\delta) < \tau$

This optimization is done in a two step procedure:

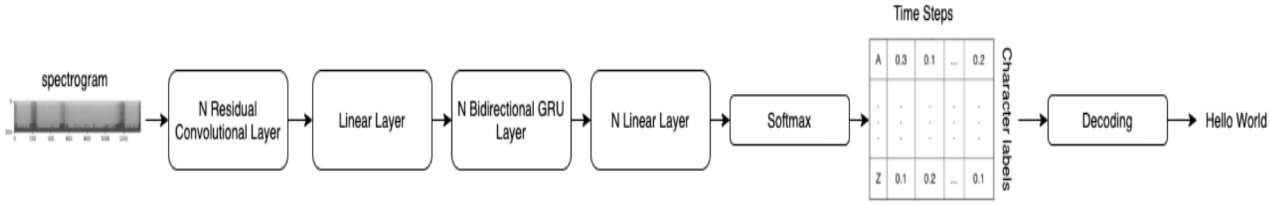
1. First we fix a threshold τ and generate an adversarial example by minimizing the CTC loss. We terminate gradient descent when the example is classified as the target.
2. We then reduce the threshold τ and attempt to generate an example for the new threshold.

This two step procedure is needed because optimization directly using l_∞ metric will often oscillate around the optimal solution without converging [1].

4 Implementation

4.1 Training a Speech to Text Model

The brief architecture of the Speech to Text Model which we used is as follows:



It is a BiRNN based architecture that uses CTC-loss.

Dataset used for Training : SpeechCommands (subset)
Number of Training examples : 12000
Number of Test examples : 4000
Number of Distinct Labels : 35
Length of Audio Clips : 1 second

The model was trained for 20 epochs using Adam optimizer with an initial learning rate of 1e-3. We had GPU access via Google Colab. The resultant model having WER of 0.16 was saved and used for further for adversarial example generation.

4.2 Performing adversarial attack

Adversarial example generation procedure :

1. A sample example which we will refer as 'original example' is selected from the dataset to generate it's adversarial version.
2. A zero vector referred as δ of the size same as the input waveform is initialised.
3. δ is clamped between $(-\tau, \tau)$ where tau is a fixed constant and rescaled by the factor of $rescale$
4. This modified δ is added to the 'original example' and passed through the Model and using the loss function as specified above gradients are backpropagated and only the value of δ is updated.
5. If the decoded prediction of the model is equal to the target label, then the $rescale$ value is reduced by a factor and the current of δ is tracked.
6. Steps 3 to 5 are repeated for a large fixed number of iterations to arrive at a best possible solution of δ

The δ obtained from the above procedure is finally added to the 'original example' to give the Adversarial example

4.3 Hyperparameter Tuning

The hyper parameters for generating adversarial examples are : τ, $learning\_rate$, $rescale$ We set τ (initial perturbation magnitude) to be 2000, learning rate as 1, and rescale as 0.9 (factor by which the max perturbation magnitude is reduced.

5 Evaluation

Metric used for Evaluation : Signal to Noise Ratio (SNR)

SNR is defined in the following way :

$$SNR(in\ dB) = 10 * log_{10}\left(\frac{||clean\ audio||_2^2}{||adversarial\ audio - clean\ audio||_2^2}\right)$$

The higher is the value of SNR, the better is our Adversarial example given that the model outputs the target label on passing this adversarial example.

We generated a total of 14 adversarial examples for target labels out of the dataset from which the model was trained. The table below shows the evaluation of each example:

	Original Audio			**Adversarial Audio**			**SNR**
	Label	Prediction	Audio (in dB)	Target Label	Target in data	Noise (in dB)	
1	sheila	sheila	19.16	hello	No	2.18	16.97
2	forward	forward	18.09	protect	No	1.04	17.05
3	backward	backward	23.66	speech	No	-9.92	33.59
4	forward	birani	19.19	predict	No	47.13	-27.94
5	nine	nine	16.7	viva	No	-22.07	38.78
6	down	down	20.39	lazy	No	-17.78	38.18
7	happy	happy	8.5	crazy	No	-23.95	32.46
8	sheila	sheila	4.15	igloo	No	-27.65	31.81
9	backward	backward	21.52	software	No	-5.74	27.27
10	seven	sevn	15.53	happy	Yes	-8.28	23.82
11	three	three	20.91	left	Yes	-20.68	41.59
12	marvin	marvin	19.95	learn	Yes	-15.82	35.78
13	forward	forward	14.11	marvin	Yes	-28.16	42.27
14	sheila	sheila	22.43	visual	Yes	-28.19	50.63

Link to the Audio files : Audio files

As seen in the above table, we achieved a reasonably good SNR values for all the examples except one (which we will mention ahead). A very little noise (in some examples no noise is heard in the adversarial audio) is heard by the human ear in almost all the examples having high SNR values. We find there is no significant difference based on whether the target is present in the dataset or not.

Example no. 4 in the above table has a negative SNR because the original audio is corrupted. It can be considered as a random audio having no meaningful content. Hence, for model to predict the target label, a high amount of perturbation is required which leads to a negative SNR value. This may indicate that it is more difficult to transform noise to a target as compared to a meaningful utterance.

6 Conclusion

During the project we gained a good understanding of adversarial attacks in the audio domain. Though we were not able to implement as much as we planned, through an extensive survey of papers we were able to understand the problem well. We successfully generated adversarial examples for a model trained on the SpeechCommands dataset. We were able to achieve a good Signal to Noise Ratio (no comparison since SNR was not reported in [1] and [2]) for most of the targets. Our results were relatively more imperceptible as compared to [1] on certain targets. This could be due to two reasons: i) our targets were single words as compared to phrases used in [1], ii) some of the targets were already present in the traning dataset.

7 Future Work

One interesting direction that we wanted to explore was adversarial attacks on Indian and other low resource languages. As these attacks are not data driven, it should be possible to directly apply them to different languages. However the success rate would be limited by the quality of the ASR systems. Also there are few reliable defenses against Adversarial Audio Attacks, and several have been shown to be ineffective. It would be interesting to work on developing effective, robust defense techniques.

References

[1] Carlini and Wagner. Audio Adversarial Examples: Targeted Attacks on Speech-to-Text https://people.eecs.berkeley.edu/~daw/papers/audio-dls18.pdf

[2] Qin et Al. Imperceptible, Robust, and Targeted Adversarial Examples for Automatic Speech Recognition https://nicholas.carlini.com/papers/2019_icml_imperceptibleaudio.pdf

[3] Abdoli et Al. Universal Adversarial Audio Perturbations https://arxiv.org/pdf/1908.03173.pdf

[4] Neekhara et Al. Universal Adversarial Perturbations for Speech Recognition Systems http://cseweb.ucsd.edu/~jmcauley/pdfs/interspeech19b.pdf

[5] Mendes and Hogan. Defending Against Imperceptible Audio Adversarial Examples Using Proportional Additive Gaussian Noise https://math.mit.edu/research/highschool/primes/materials/2020/Mendes-Hogan.pdf

[6] Yang et Al. Characterizing Audio Adversarial Examples Using Temporal Dependency https://openreview.net/pdf?id=r1g4E3C9t7

[7] Taori et Al. Targeted Adversarial Examples for Black Box Audio Systems https://arxiv.org/pdf/1805.07820.pdf

[8] Taori et Al. Intriguing Properties of Neural Networks https://arxiv.org/pdf/1312.6199.pdf