Towards Evaluating Adversarial Attacks in Audio Domain

Yigit Alparslan
Drexel University
Philadelphia, US
ya332@drexel.edu

ABSTRACT

This paper investigates adversarial attacks on neural networks in audio domain. Adversarial attacks are inputs that look like the original input but altered on purpose. Speech-to-text neural networks that are widely used today are prone to misclassify adversarial attacks[1][2][3]. In this study, first, we investigate the presence of targeted adversarial attacks[4][11] by altering wave forms from Common Voice data set. We craft adversarial wave forms via Connectionist Temporal Classification Loss Function[13][14][15] and attack DeepSpeech[12]– speech-to-text neural network implemented by Mozilla. We achieve 100% adversarial success rate(0 successful classification by DeepSpeech) on all 25 adversarial wave forms that we crafted. Second, we investigate the use of PCA as a defense mechanism against adversarial attacks. To smooth out the adversarial noise, we reduce dimensionality by applying PCA to these 25 attacks that we created, and test them with DeepSpeech again. We achieve 100% adversarial success again, which suggests PCA is not a good defense mechanism in audio domain. Finally, instead of using PCA as a defense mechanism, we use PCA this time to craft adversarial inputs under a black-box setting with minimal adversarial knowledge. With no knowledge regarding the model, parameters, or weights, we craft another 25 adversarial attacks by applying PCA to samples from Common Voice data set, and achieve 100% adversarial success under black-box setting again when tested against DeepSpeech. We also experiment with different percentage of components necessary to result in a classification during attacking process. In all cases, adversary becomes successful. However, in the case of 95% reconstruction, we achieve average similarity ratio of 87.85% percent and normalized edit distance of 9 (adversarial input is very similar to the original), whereas in the case of 10% reconstruction, we achieve average similarity ratio of 27.42%, and normalized edit distance of 46(adversarial input sounds corrupted to human ear).

Author Keywords

Audio attacks, DeepSpeech, Adversarial Attacks, Speech-to-Text Neural Network, Connectionist Temporal Classification, Recurrent Neural Networks

INTRODUCTION

Numerous recent studies have demonstrated how Deep Neural Network (DNN) classifiers can be fooled by adversarial examples[1],[2],[3],[4],[5], in which an attacker adds perturbations to an original sample, causing the classifier to misclassify the sample [6]. Adversarial attacks that render DNNs vulnerable in real life represent a serious threat, given the consequences of improperly functioning autonomous vehicles, malware filters, or biometric authentication systems [7][8][9]. Social network companies, law enforcement, and various commercial interests may wish to be able to use image classification and speech-to-text transcription tools with defensive mechanisms that are robust to adversarial attacks[2][11], in order to reduce these attacks effectiveness. Wide usage of DNNs makes the problem of creating robust and secure DNNs even more important in safety-critical applications.

We are interested in exploring whether there are any particular characteristics of audio recognition which seem to make adversarial attacks more or less successful in this area. Current attacks that have been studied were primarily in the image domain[4][5][6][7], such as those of Carlini [2] and Papernot et al. [17] Such attacks have been studied as proof-of-concepts, where adversarial attackers are assumed to have full knowledge of the classifier (e.g. model, architecture, model weights, parameters, training and testing data sets). The strongest attack in the literature at the time of writing this article is Carlini's attack [11] based on the L_2 norm, and it is a white-box attack requiring full knowledge of the model. Much of this research has been interested in developing the most effective attacks possible, to be used as standards against which to test the robustness of classifier RNNs[2]. With less knowledge of the classifier model, the effectiveness of the attack decreases. There is also interest in crafting attacks that assume minimal knowledge of the adversary–that is, attacks under black-box setting – regarding the classifier model, since in most real-world applications the adversary does not have access to the classifier's parameters unless the adversary is an insider. We are inspired by work on image domain[2][3][4][5][6], and investigate the audio domain in this paper.

We use Common Voice Data set, which is widely used as a standard audio data set in the literature[11]. Being a common data set, it provides a common ground for different research