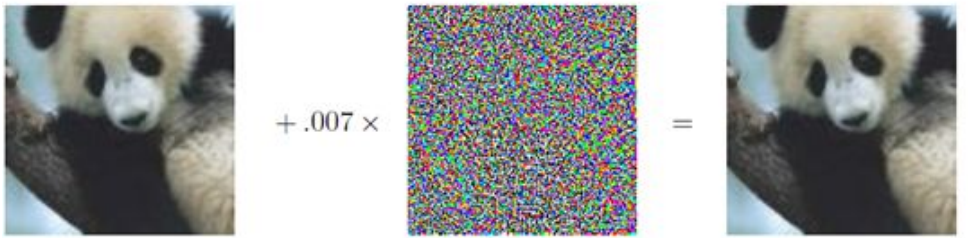
Targeted Speech Adversarial Example Generation With Generative Adversarial Network

DONGHUA WANG, LI DONG, RANGDING WANG, DIQUN YAN AND JIE WANG

Presented By: Hiskias Dingeto

INTRODUCTION

- Adversarial Examples
  - Explaining and Harnessing Adversarial Examples, 2015
- Modifying the features of an instance of inputs intentionally in order to cause misclassification



INTRODUCTION

- Adversarial Examples in The Physical World
  - Apply slight modifications to real life images and causing misclassification
  - Implementing Adversarial Attacks on roadside traffic signs

INTRODUCTION

- Automatic Speech Recognition Systems
  - Apply Siri
  - Amazon Alexa
  - Google Assistance
- Various systems aim to attack ASRs in different ways
- Other systems as compared to this system one of the following two drawbacks
  - Large computational cost
  - Perturbation added to the adversarial examples noticeably lowers the quality

INTRODUCTION

- Attacking neural network based speech recognition models (Automatic Speech Recognition Systems or ASRs)
- Using GANs for constructing targeted speech adversarial examples
- Goal of Generator:
  - Generating noise that can cause misclassification
  - Fooling discriminator from distinguishing adversarial example from a normal sample
- Goal of Discriminator:
  - Distinguish generated adversarial examples from normal samples

RELATED WORK

- Generative Adversarial Network
  - Min-Max Game
  - Realistic image generation, image to image translation, text-to-image synthesis, adversarial example generation
- Speech Adversarial Example Generation
  - Fast Gradient Sign Method (FGSM)

PROBLEM FORMULATION

- Two types of Adversarial Attacks
  - Targeted Attacks: cause the system to misclassify to an known label
  - Untargeted Attacks: cause the system to misclassify to an unknown label hence untargeted
- Targeted Attacks are more difficult to implement
- Goal of the system:
  - Generate an adversarial example that could fool the ASR
  - Adversarial example should be similar to the original sample
- Use GANs to generate perturbation

TARGETED SPEECH ADVERSARIAL EXAMPLE GENERATION USING GAN

- Framework overview:

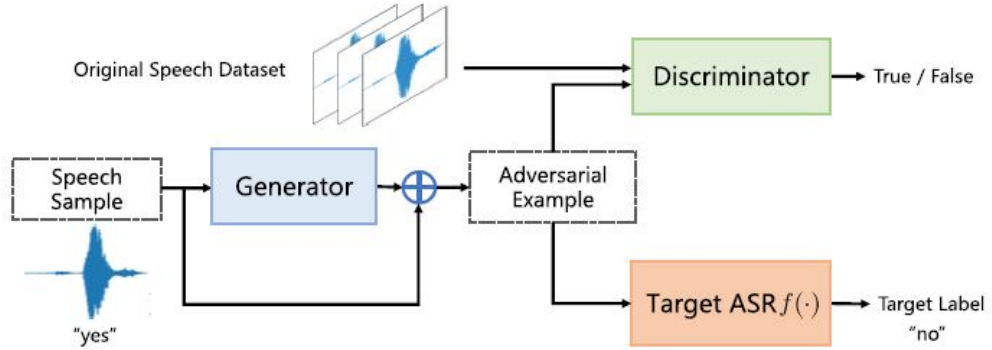


FIGURE 1. The overview of the proposed framework.

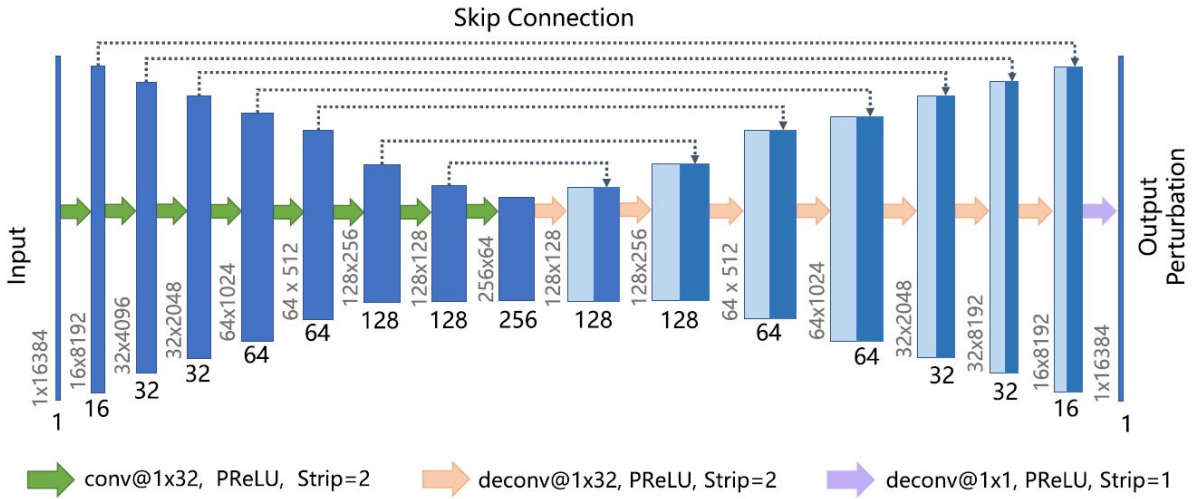
TARGETED SPEECH ADVERSARIAL EXAMPLE GENERATION USING GAN

- Generator
  - Encoder decoder like network (UNet)
  - (explained in the next figure)
- Discriminator
  - (explained in the next figure)
- Loss function
  - Target Classifier Loss
  - Adversarial Loss
  - Regularization terms

$$L_{\mathcal{G}} = L_{\text{adv}}^{f} + \alpha L_{\text{fool}} + \beta L_{\text{hinge}} + \gamma L_2,$$

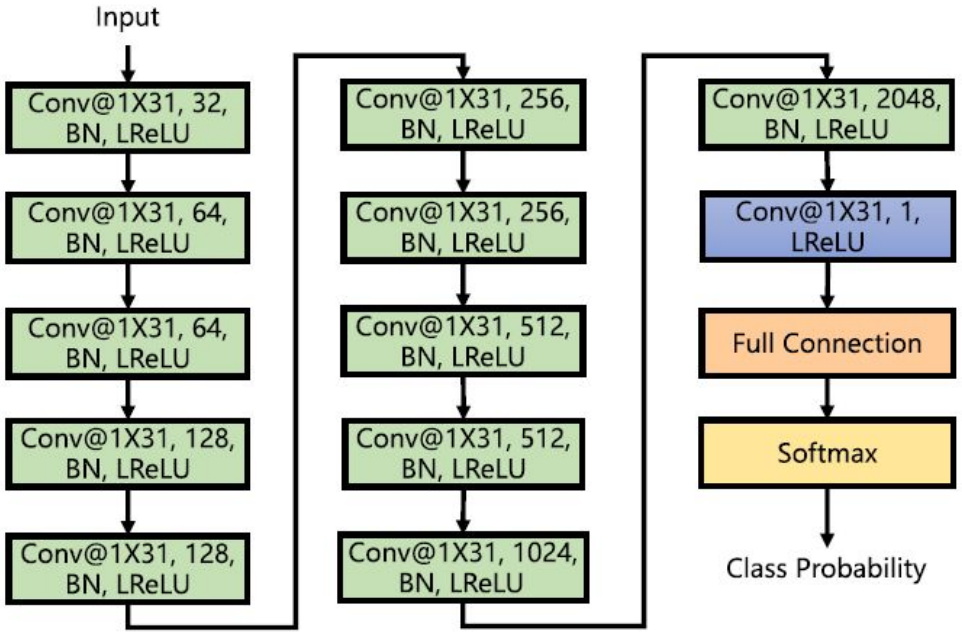
TARGETED SPEECH ADVERSARIAL EXAMPLE GENERATION USING GAN

- Architecture of the Generator



TARGETED SPEECH ADVERSARIAL EXAMPLE GENERATION USING GAN

- Architecture of the Discriminator



EXPERIMENTAL RESULTS

- Performance Metric 1: Success Rate
  - success_rate = #{misclassified_samples} / #{test_samples}
- Performance Metric 2: Objective Quality
  - Signal-to-Noise Ratio (SNR)
  - PESQ

$$\mathrm{SNR}(x^{\mathrm{adv}}) = 10 \cdot \log_{10} \frac{P_x}{P_\delta},$$

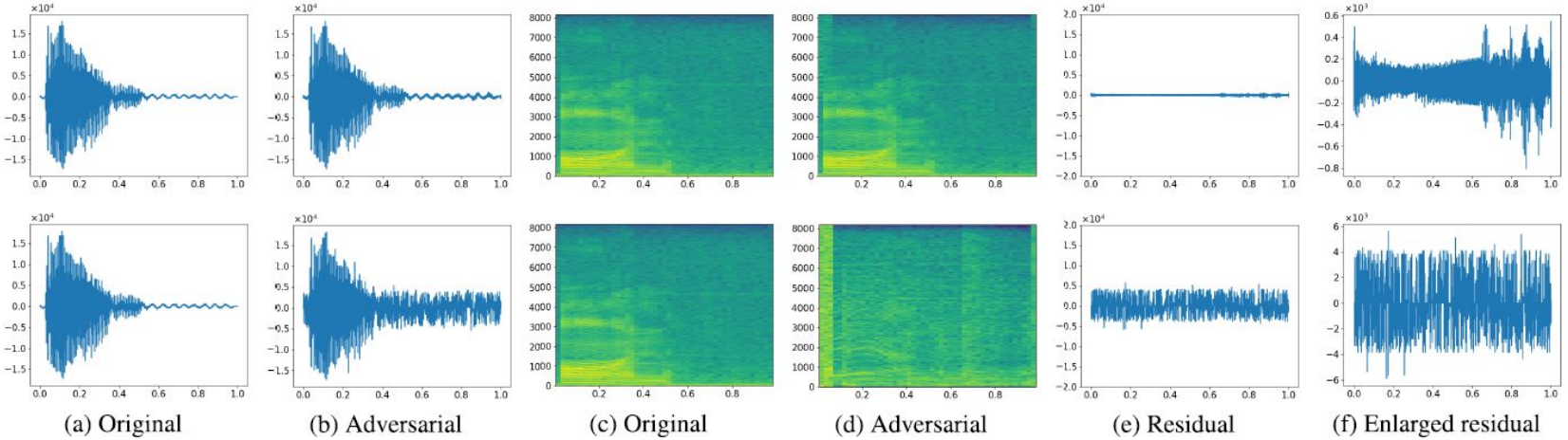
EXPERIMENTAL RESULTS

- Comparison of the proposed method with Alzantot and SirenAttack

Attacking Method	Attacking WideResNet on SpeechCmd			Attacking SampleCNN on GTZAN		
	success_rate	SNR(dB)	Time(s)	success_rate	SNR(dB)	Time(s)
Alzantot *et al.* [9]	84.96%	15.72	231.46	82.76%	14.15	215.36
SirenAttack [10]	89.25%	17.57	368.29	89.10%	15.39	452.21
Proposed	**92.33%**	**20.27**	**0.009**	**90.58%**	**26.92**	**0.01**

EXPERIMENTAL RESULTS

- Comparison between original speech sample with generated adversarial sample



FUTURE PLAN

- Research more on Adversarial Attacks
- Counteracting their effects in Machine Learning Models
- Find out more about what causes Machine Learning Models to misclassify Adversarial Examples