TẤN CÔNG HỆ THỐNG NHẬN DIỆN GIỌNG NÓI BẰNG CÁCH TẠO MẪU ĐỐI KHÁNG CÓ MỤC TIÊU

Nguyễn Trường Thịnh[1,2]

[1]21520110@gm.uit.edu.vn

[3]Trường Đại học Công nghệ Thông tin ĐHQG TP.HCM

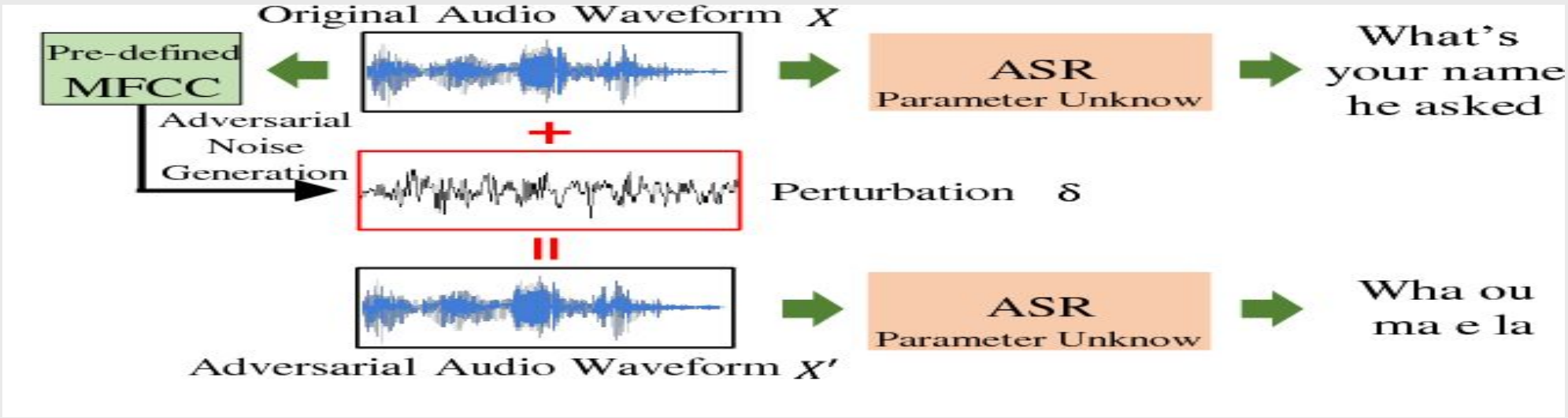
Tóm tắt

Mô hình nhận diện giọng nói tự động (ASR) là một thành quả của trí tuệ nhân tạo (AI) ngày nay. Việc tấn công các hệ thống này là một bài toán cần được nghiên cứu và bận tâm. Chúng tôi giới thiệu một bài toán sử dụng mẫu đối kháng adversarial example attacks để đánh lừa các mô hình học sâu trong ASR sinh ra một bản dịch (transcript) đã được soạn sẵn.
Trong đó đề xuất một hàm mất mát CTC-loss và hai quy trình tạo mẫu đối kháng trên ngữ cảnh white-box

Động lực

Ngày nay các ứng dụng của mô hình nhận diện giọng nói tự động nổi bật nhất như: các trợ lý ảo (Apple Siri, Google Assistant hay Amazon Alexa), điều khiển xe tự hành, phiên dịch được áp dụng rộng rãi sử dụng deep learning.
Việc tấn công bất kì lỗ hổng nào của mô hình deep learning đều gây ảnh hưởng lớn, làm tê liệt toàn bộ hệ thống.
Chúng ta cần phải nhận thức được lỗ hổng, khả năng bị tấn công và các phương pháp phòng thủ để bảo vệ hệ thống ASR, và cho chính chúng ta.

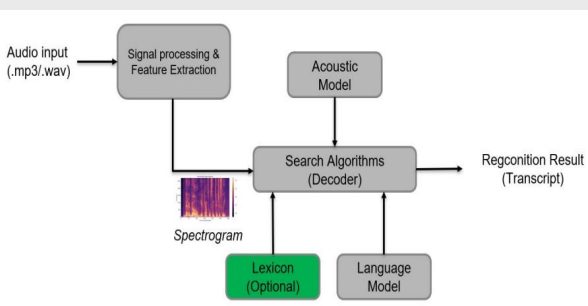
Tổng quan



Mô tả

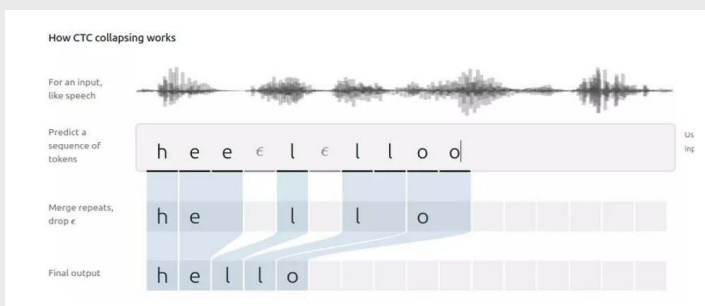
1. Mô hình ASR:

một hệ thống ASR gồm 3 phần riêng biệt là: acoustic model, lexicon, language model, sau đó dùng thuật toán tìm kiếm như một bộ decoder. Tuy nhiên, sự phát triển mạnh mẽ của các mô hình học sâu đã đề xuất mô hình end-to-end đó là CTC(Connectionist Temporal Classification) chịu trách nhiệm cho gần như toàn bộ quá trình chuyển đổi (decoding) các đặc trưng thành một bản dịch hoàn chỉnh.



2. Cách hoạt động của CTC

Trong ASR, đầu ra thường sẽ là một câu chưa hoàn chỉnh vì có các kí tự lặp lại như "ccaaattt", "doogg",...Nguyên nhân dẫn tới những hiện tượng này là do giọng nói dài (giọng ngân dài khi nói, ngân nga khi hát,...), giọng bị ngắt quãng,... Do đó, để cho ra được một câu hoàn chỉnh thì ta cần phải căn chỉnh đầu ra, loại bỏ các ký tự lặp lại được giải quyết bằng CTC. Cách CTC hoạt động là tạo ra một số blank ngăn cách sau đó co các ký tự giống nhau lại tạo ra nhiều alignment khác nhau (như ảnh dưới). Xác suất dự đoán sẽ bằng tổng xác suất của các alignment.



3. Tạo mẫu đối kháng:

Quy trình tạo mẫu đối kháng gồm 2 giai đoạn:
Giai đoạn 1: sẽ tập trung xây dựng mẫu đối kháng để đánh lừa mô hình bằng phương pháp của Carlini & Wagner.
Giai đoạn 2: thực hiện áp dụng psychoacoustic model để giúp mẫu khó có thể nhận ra bởi tai người.

4. Đánh giá & kết quả mong đợi:

Thực hiện tấn công hai mô hình ASR là DeepSpeech và Lingvo với mã nguồn mở, trên tập dataset LibriSpeech.
Mong muốn mô hình phải có tỉ lệ mẫu đối kháng được chuyển thành bản dịch theo mục tiêu của attacker trên 85%.
Có thể đưa vào ứng dụng thực tế trong môi trường tự nhiên có tồn tại nhiễu.