NHẬN DẠNG GIỌNG NÓI TIẾNG VIỆT TỰ ĐỘNG

Nguyễn Đặng Quang Tuấn [1,3] **Hồ Thanh Tịnh** [2,3]

{[1]20522116, [2]20520813}@gm.uit.edu.vn

[3] Trường Đại học Công Nghệ Thông Tin ĐHQG TP.HCM

Đề tài?

Nghiên cứu mô hình nhận diện giọng nói Wav2vec 2.0

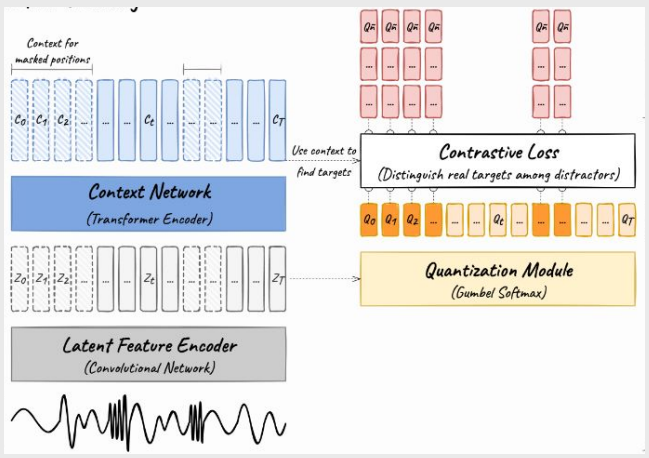
- Tìm hiểu các kĩ thuật tạo nên sự ưu việt của mô hình
- Huấn luyện mô hình với dataset Tiếng Việt
- Xây dựng hệ thống điều khiển bằng giọng nói thử nghiệm.

Lý do chọn đề tài?

- Các hệ thống điều khiển bằng giọng nói rất hữu ích và ngày càng được áp dụng nhiều trong đời sống.
- Cần giải quyết bài toán nhận dạng giọng nói tự động (ASR).
- Học có giám sát là giải pháp phổ biến cho ASR, tuy nhiên cần lượng lớn dữ liệu có gán nhãn để đạt được hiệu suất chấp nhận được.

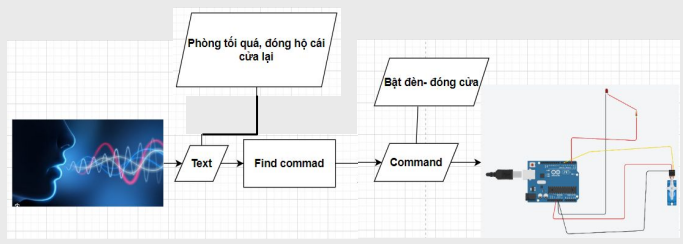
Tổng quan

Wav2vec 2.0 → **Tinh chỉnh và đánh giá** → **Hệ thống điều khiển bằng giọng nói**



VLSP

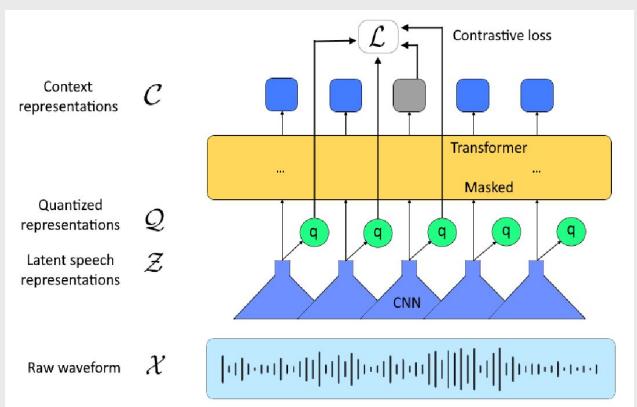
WER



Mô tả

1. Vì sao lại là Wav2vec 2.0

- Lượng dữ liệu giọng nói có gán nhãn đang có khá nhỏ. Lượng dữ liệu không nhãn thì rất lớn nhưng chưa được tận dụng.
- Tham khảo tại các hội nghị CVPR, ICCV, ECCV, NIPS.
  Các nguồn: paperwithcode, google scholar,...
  Nhận thấy Wav2vec 2 là giải pháp thích hợp cho bài toán ASR



- Self supervised learning: Chỉ cần tinh chỉnh với lượng nhỏ labeled data mô hình có thể có hiệu suất đánh bại các mô hình supervised hiện đại nhất.

Main idea:

- Wav2vec 2.0 được huấn luyện theo hai giai đoạn.
- Giai đoạn 1: mô hình học ra các cấu trúc ẩn bên trong lượng lớn unlabeled data
- Giai đoạn 2: tinh chỉnh có giám sát với labeled data.
- Contrastive learning: Tương tự như word embedding, mô hình học cách biểu diễn các giọng nói gần nhau bằng các vector gần nhau từ unlabeled data.
- Các kĩ thuật: CNN, Transformer, Quantization module, Masking, Gumbel softmax.

2. Huấn luyện và đánh giá

- Sử dụng model pretrained trên 13k giờ unlabeled data từ youtube cho giai đoạn 1.
- Fine tune trên 250h labeled data VLSP-ASR dataset ở giai đoạn 2.
- Fine tune bổ sung trên bộ dữ liệu BKAI2023 SLU.
- Mô hình được đánh giá bằng thang đo WER.

3. Hệ thống thử nghiệm

- Hệ thống điều khiển 2 thiết bị led và servo.
- Mô hình ngôn ngữ được sử dụng để hiểu được câu lệnh điều khiển theo ngữ cảnh của câu nói từ người dùng.
- Ví dụ: "sao phòng tối om vậy, tôi muốn ánh sáng và hãy mở cửa ra giúp tôi" câu lệnh điều khiển sẽ là "bật đèn" , "mở cửa".

4. Kết quả mong đợi

- Hiểu được các kĩ thuật được sử dụng trong bài báo gốc.
- Mô hình được tinh chỉnh trên dữ liệu Tiếng Việt sẽ đưa ra các dự đoán có độ chính xác cao được đánh giá bởi thang đo WER.
- Xây dựng được hệ thống thử nghiệm hoạt động đúng như mong đợi. Có thể điều khiển các thiết bị cho trước bằng giọng nói đúng yêu cầu.

NII

Nguyễn Đặng Quang Tuấn– Trường Đại học

TEL : 0919895210 Email : 20522116@gm.uit.edu.vn