CẢI THIỆN WAV2VEC2.0 TRONG NHẬN DẠNG TIẾNG NÓI THÔNG QUA NOISY STUDENT TRAINING VÀ CẤP PHỤ TỪ

Hoàng Anh Đức Đăng Quang [1,2] **Phan Văn Thiện** [1,2]

[1] Trường Đại học Công nghệ thông tin, ĐHQG-HCM

[2] Khoa học máy tính

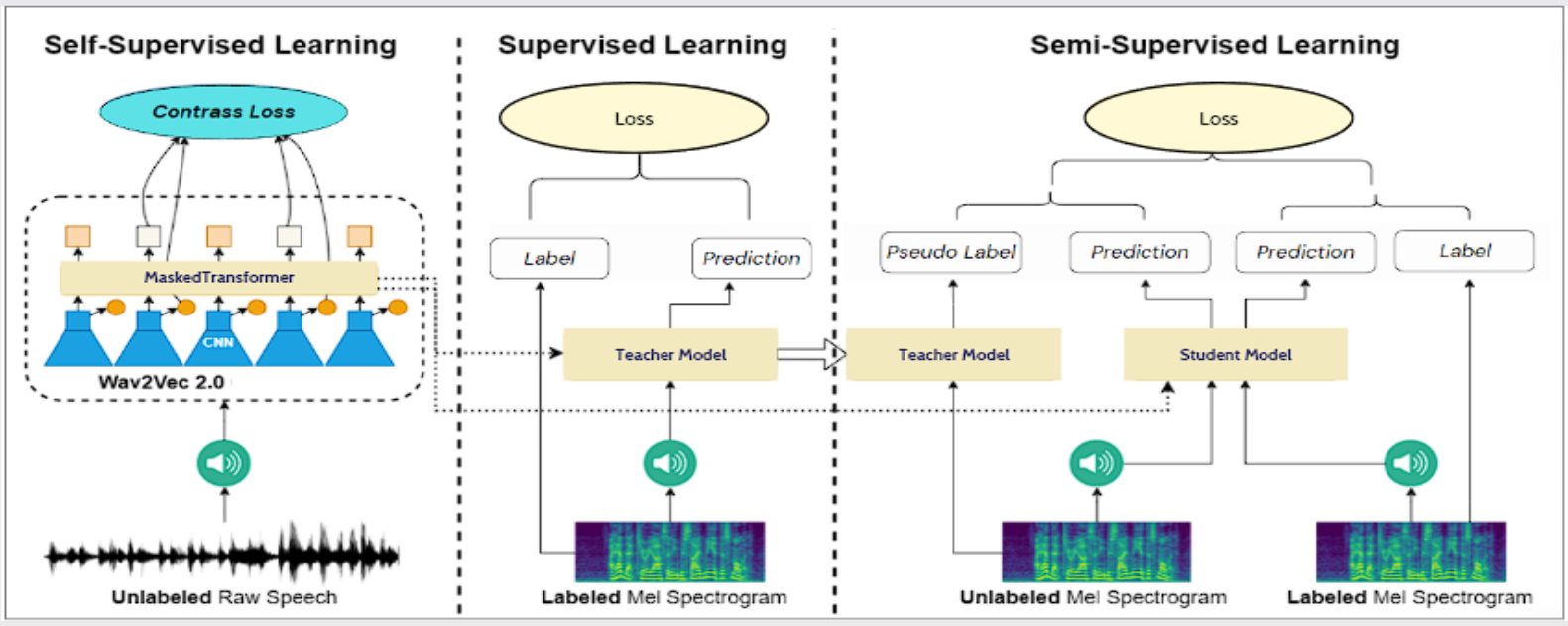
Mục tiêu

- Nghiên cứu mô hình **Wav2vec2.0** hiện có và cải thiện hiệu suất của **Wav2vec2.0** trong bài toán Nhận dạng tiếng nói tự động thông qua tinh chỉnh cấp độ phụ từ.
- Áp dụng khung huấn luyện **Noisy Student Training** cho mô hình **Wav2vec2.0** nhằm cải thiện độ chính xác và tận dụng nguồn dữ liệu không gán nhãn.
- Huấn luyện tinh chỉnh mô hình **Wav2vec2.0** trên khung huấn luyện cải tiến cho bộ dữ liệu **VLSP2021** cho bài toán **Nhận dạng tiếng nói tiếng Việt**.

Lý do chọn đề tài ?

- Nhận dạng tiếng nói là một bài toán quan trọng, hướng tới việc chuyển đổi âm thanh thành văn bản.
- Mô hình **Wav2vec2.0**, một trong những phương pháp tiên tiến hiện nay, đã đạt được hiệu suất cao. Tuy nhiên, nó vẫn có một số điểm chưa tối ưu như không tận dụng tối đa khả năng ở cấp độ ký tự và chưa có giải pháp tận dụng nguồn dữ liệu không gán nhãn sẵn có.

Overview



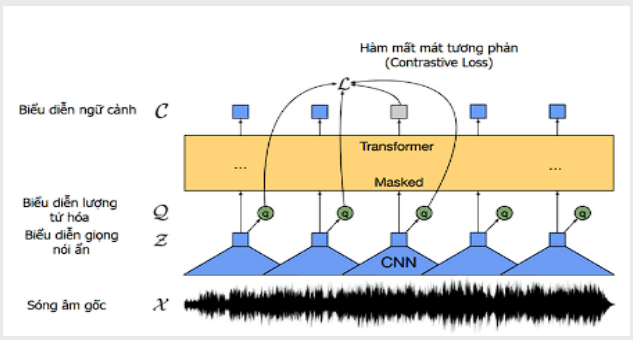
Description

1. Nội dung

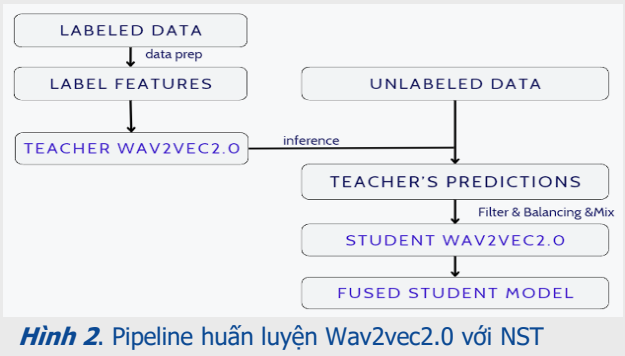
- Nghiên cứu mô hình **Wav2vec2.0** trong bài toán Nhận diện tiếng nói tiếng Việt.
- Nghiên cứu kỹ thuật Noisy Student Training, so sánh, đánh giá tính khả thi và điều chỉnh các tham số phù hợp nhất cho mô hình **Wav2vec2.0**.
- Nghiên cứu tác động của các cấp độ từ đến mô hình trong bài toán Nhận diện giọng nói.
- Huấn luyện mô hình trên bộ dữ liệu **VLSP2021** để đánh giá hiệu suất của mô hình với kỹ thuật huấn luyện mới so với việc huấn luyện cơ bản

2. Phương pháp

- Tìm hiểu kiến trúc mô hình Wav2vec2.0 trong bài toán Nhận diện tiếng nói, sự khác biệt khi sử dụng hàm mục tiêu **CTC** và **RNNT** trong **Wav2vec2.0**, những thay đổi của mô hình khi huấn luyện trên dữ liệu âm thanh tiếng Việt.
- Tìm hiểu kỹ thuật Noisy Student Training dựa theo một số kết quả đã được thực nghiệm trước đó. Thực hiện thử nghiệm Wav2vec2.0 với NST và thực hiện điều các tham số tỉ lệ trộn, số thế hệ để tìm được các tham số phù hợp.
- Tìm hiểu về sự khác biệt giữa cấp độ ký tự và cấp độ phụ từ, bao gồm yêu cầu tài nguyên huấn luyện, lượng dữ liệu và ảnh hưởng đối với tỉ lệ lỗi và độ chính xác thông qua các nghiên cứu và thử nghiệm.
- Huấn luyện mô hình Wav2vec2.0 đã được sửa đổi cấp độ từ trên nhiều thế hệ **Noisy Student Training** để đánh giá giới hạn của mô hình trên kỹ thuật huấn luyện mới. Thực hiện huấn luyện trên bộ dữ liệu âm thanh tiếng Việt VLSP2021, so sánh và đánh giá dựa trên tỉ lệ lỗi từ (WER)



Hình 1. Cấu trúc Wav2vec2.0



Hình 2. Pipeline huấn luyện Wav2vec2.0 với NST

3. Kết quả mong đợi

- Xây dựng thành công hệ thống nhận dạng giọng nói tiếng Việt với kết quả tốt hơn mô hình tiêu chuẩn và đạt WER kỳ vọng dưới 6% trên benmark VIVOS và dưới 10% trên benmark Commonvoice Vi.

Hoàng Anh Đức Đăng Quang, Phan Văn Thiện – Trường Đại học Công nghệ thông tin, ĐHQG-HCM
Email : 21522509@gm.uit.edu.vn 21522628@gm.uit.edu.vn