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

Hoàng Anh Đức Đăng Quang- 21522509

Phan Văn Thiện- 21522628

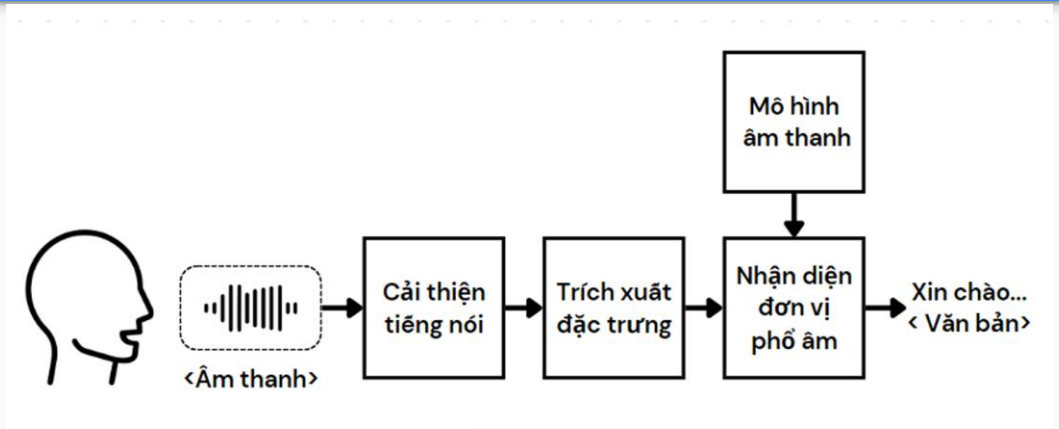
Tóm tắt

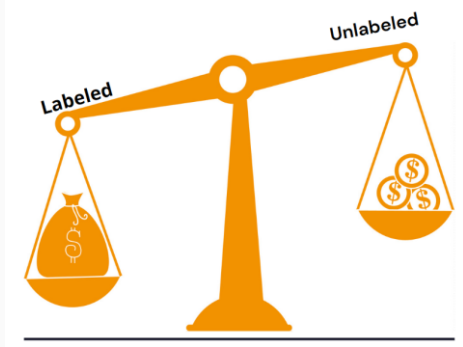
- Lớp: CS519.O11
- Link Github của nhóm: https://github.com/QuangHoang059/CS519.O11
- Link YouTube video: https://youtu.be/tHqEe6ect68

Hoàng Anh Đức Đăng Quang

Phan Văn Thiện

Giới thiệu





SOS	S	I	N	H	_	V	I	Ê	N	EOS

Cấp độ ký tự

SOS	SINH	_	VIÊN	EOS

Cấp độ phụ từ

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

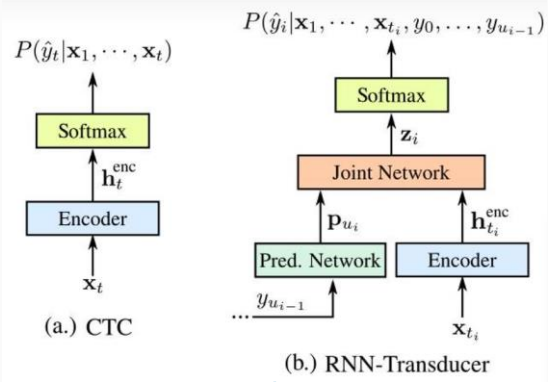
- Áp dụng khung huấn luyện Noisy Student Training cho mô hình Wav2vec2.0 nhằm cải thiện độ chính xác và tận dụng nguồn dữ liệu không gán nhãn

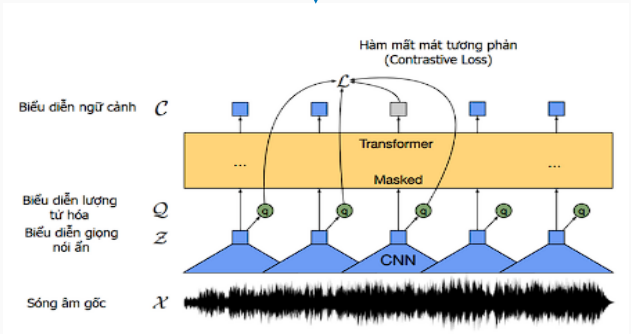
- Huấn luyện tinh chỉnh mô hình Wav2vec2.0 trên khung huấn luyện cải tiến cho bộ dữ liệu VLSP2021 cho bài toán Nhận dạng tiếng nói tiếng Việt

Nội dung và Phương pháp

1. Nghiên cứu mô hình Wav2vec2.0.
2. Nghiên cứu tác động của cấp độ từ và tìm hàm mục tiêu phù hợp.

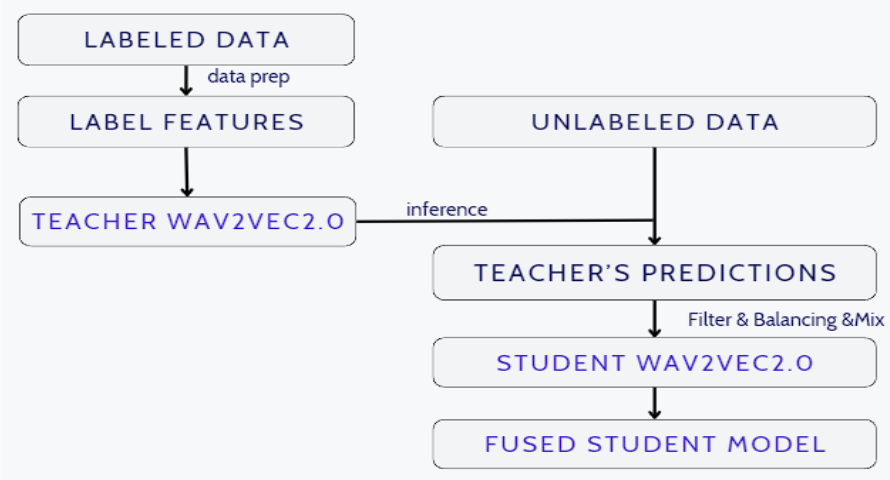
- Tìm hiểu sự khác biệt giữa các cấp độ từ.
- Tìm hiểu hàm mục tiêu CTC và RNNT





Nội dung và Phương pháp

3. Nghiên cứu kỹ thuật NST áp dụng trên mô hình Wav2vec2.0.
   - Tìm hiểu và điều chỉnh các tham số phù hợp.
4. Huấn luyện với khung huấn luyện mới và đánh giá.
   - Huấn luyện trên bộ dữ liệu VLSP2021
   - Đánh giá trên Tỉ lệ lỗi từ (WER)



Kết quả dự kiến

- Báo cáo các phương pháp và kỹ thuật của khung huấn luyện mới dành cho Wav2vec2.0 được xây dựng và phát triển cho bài toán nhận diện tiếng nói: Kết quả thực nghiệm, đánh giá, so sánh so với phương pháp ban đầu.
- 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.

Tài liệu tham khảo

[1]Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, Michael Auli: wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *NeurIPS* (2020)
[2] Qizhe Xie, Minh-Thang Luong, Eduard H. Hovy, Quoc V. Le: Self-Training
With Noisy Student Improves ImageNet Classification. CVPR 2020: 10684-10695
[3] Yu Zhang, James Qin, Daniel S. Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc V. Le, Yonghui Wu: Pushing the Limits of Semi-Supervised Learning for Automatic Speech Recognition. CoRR abs/2010.10504 (2020)
[4] Jan Kremer, Lasse Borgholt, Lars Maaløe: On the Inductive Bias of Word-Character-Level Multi-Task Learning for Speech Recognition. CoRR abs/1812.02308 (2018)
[5] Yu Zhang, James Qin, Daniel S. Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc V. Le, Yonghui Wu: Pushing the Limits of Semi-Supervised Learning for Automatic Speech Recognition. CoRR abs/2010.10504 (2020)
[6] Alex Graves:Sequence Transduction with Recurrent Neural Networks. CoRR abs/1211.3711 (2012)
[7] Albert Zeyer, Kazuki Irie, Ralf Schlüter, Hermann Ney:Improved training of end-to-end attention models for speech recognition. CoRR abs/1805.03294 (2018)
[8] Alex Graves, Santiago Fernández, Faustino J. Gomez, Jürgen Schmidhuber:Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. ICML 2006: 369-376