


THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút):
<https://youtu.be/tHqEe6ect68>
- Link slides (dạng .pdf đặt trên Github của nhóm):
https://github.com/QuangHoang059/CS519.O11/blob/main/ENHANCE_WA_V2VEC2.0_FOR_SPEECH_RECOGNITION.pdf
- Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới
- Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in

<ul style="list-style-type: none">• Họ và Tên: Phan Văn Thiện• MSSV: 21522628 	<ul style="list-style-type: none">• Lớp: CS519.O11• Tự đánh giá (điểm tổng kết môn): 9/10• Số buổi vắng: 0• Số câu hỏi QT cá nhân: 11/11• Link Github: https://github.com/QuangHoang059/CS519.O11/• Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:<ul style="list-style-type: none">◦ Viết tài liệu đề cương◦ Làm video thuyết trình Youtube
<ul style="list-style-type: none">• Họ và Tên: Hoàng Anh Đức Đăng Quang• MSSV: 21522509 	<ul style="list-style-type: none">• Lớp: CS519.O11• Tự đánh giá (điểm tổng kết môn): 9/10• Số buổi vắng: 1• Số câu hỏi QT cá nhân: 10/11• Link Github: https://github.com/QuangHoang059/CS519.O11/• Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:<ul style="list-style-type: none">◦ Làm slide thuyết trình◦ Làm poster

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

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

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

ENHANCE WAV2VEC2.0 FOR SPEECH RECOGNITION VIA NOISY STUDENT TRAINING AND SUBWORD LEVEL

TÓM TẮT

Nhận dạng tiếng nói là một bài toán quan trọng và có tính ứng dụng cao với mục tiêu là chuyển đổi một đoạn âm thanh thành một biểu diễn văn bản tương ứng. Cách tiếp cận phổ biến cho bài toán này hiện nay là các mô hình tiền huấn luyện, nổi bật trong số đó là mô hình Wav2vec2.0. Tuy đạt được hiệu suất cao nhưng Wav2vec2.0 vẫn còn một số vấn đề như: chưa tận dụng được tối đa khả năng của mình với cấp độ ký tự và vấn đề dữ liệu huấn luyện. Từ những vấn đề trên, chúng tôi đề xuất cải tiến mô hình như sau:

- Thứ nhất, áp dụng cấp độ phụ từ nhằm tận dụng hết khả năng của mô hình và tăng độ chính xác cho mô hình.
- Thứ hai, áp dụng khung huấn luyện Noisy Student Training vốn rất thành công ở ImageNet vào huấn luyện tinh chỉnh Wav2vec2.0 nhằm tận dụng được nguồn dữ liệu không gán nhãn một cách hiệu quả.

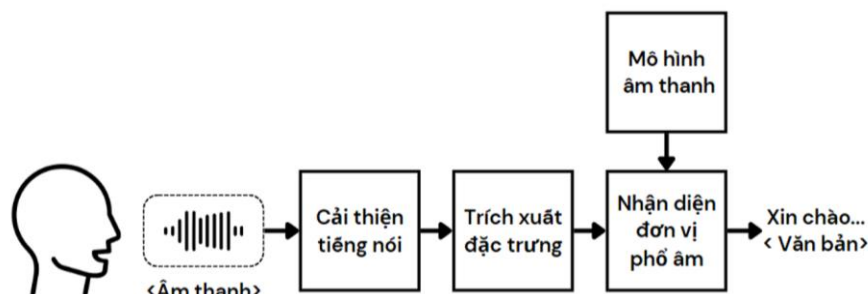
Kết quả kỳ vọng của đề tài là áp dụng thành công khung huấn luyện mới, huấn luyện trên bộ dữ liệu tiếng Việt VLSP2021 và đạt được tỉ lệ lỗi từ dưới 6% trên benchmark VIVOS và dưới 10% trên benchmark Common Voice Vi.

GIỚI THIỆU

Nhận dạng tiếng nói là một trong những bài toán quan trọng và có tính ứng dụng cao trong lĩnh vực xử lý âm thanh nói riêng và trí tuệ nhân tạo nói chung. Ứng dụng của nhận dạng tiếng nói có thể kể đến như tạo phụ đề tự động, trợ lý ảo, cải thiện giao tiếp người máy,...

Mục tiêu chính của Nhận dạng tiếng nói là chuyển đổi các âm thanh tiếng nói thành các biểu diễn văn bản. Cụ thể:

- Đầu vào: một đoạn âm thanh được thu âm từ micro hoặc tệp audio
- Đầu ra: một đoạn văn bản hoặc biểu diễn văn bản tương ứng.



Trong những năm gần đây, cách tiếp cận phổ biến để giải quyết bài toán này thường tập trung vào các mô hình tiền huấn luyện nhằm đạt độ chính xác cao hơn các cách tiếp cận dựa trên mô hình được huấn luyện từ đầu. Nổi bật trong đó là mô hình tiền huấn luyện Wav2vec2.0 [1] giúp mang lại hiệu suất cao bằng việc sử dụng kiến trúc Transformer.

Mặc dù đạt được hiệu suất cao trên bài toán nhận dạng tiếng nói, nhưng wav2vec2.0 vẫn chưa được tận dụng tối đa khả năng của mình. Hiện nay, Wav2vec2.0 chủ yếu chỉ mới dừng lại ở việc tinh chỉnh mức độ ký tự. Một số nghiên cứu [4][7] đã chỉ rằng mức độ ký tự tuy giúp tiết kiệm tài nguyên nhưng lại chưa đạt kết quả cao bằng cấp độ phụ từ, cấp độ từ. Bên cạnh đó, thiếu hụt về dữ liệu gán nhãn cũng là một vấn đề đáng quan tâm trong Nhận dạng tiếng nói nói riêng và lĩnh vực Trí tuệ nhân tạo nói chung. Để cải thiện vấn đề này, các thuật toán học bán giám sát ra đời. Trong số đó, Noisy Student Training[2] cũng là một kỹ thuật đáng chú ý và đã đạt hiệu suất tốt trong ImageNet.

Điều này đặt ra hai câu hỏi:

- Thứ nhất, làm sao để phát triển mô hình lên cấp độ phụ từ, nhằm cải thiện hiệu suất mô hình tốt hơn ?
- Thứ hai, liệu có thể để áp dụng khung huấn luyện Noisy Student Training[2] vào mô hình Wav2vec2.0, nhằm tận dụng được nguồn dữ liệu không gán nhãn phong phú, chi phí thấp?

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 và đạt được kết quả cao nhất trên benchmark VIVOS và CommonVoice Vi

NỘI DUNG VÀ PHƯƠNG PHÁP

a. NỘI DUNG:

- Nghiên cứu mô hình Wav2vec2.0 trong bài toán Nhận diện tiếng nói.
- Nghiên cứu tác động của các cấp độ từ đến mô hình trong bài toán Nhận diện giọng nói và tìm hàm mục tiêu thích hợp cho mô hình.
- 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.
- 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.

b. PHƯƠNG PHÁP:

- Tìm hiểu mô hình Wav2vec2.0 trong bài toán nhận diện tiếng nói.
- Tìm hiểu về sự khác biệt giữa các cấp độ từ: cấp độ ký tự và cấp độ phụ từ, các yêu cầu và ảnh hưởng khi chuyển đổi giữa các cấp độ từ như tài nguyên huấn luyện lượng dữ liệu tối thiểu, sự thay đổi về tỉ lệ lỗi từ và độ chính xác thông qua các bài báo [4][7] và thực nghiệm.
- Tìm hiểu hàm mục tiêu CTC [8] và RNNT[6] cho bài toán nhận diện tiếng nói, sự khác biệt về hiệu quả và cấu hình khi áp dụng với Wav2vec2.0.
- Tìm hiểu kỹ thuật Noisy Student Training dựa theo một số kết quả đã được thực nghiệm trước đó [2][3]. Thực hiện thử nghiệm Wav2vec2.0 với NST và thực hiện điều chỉnh các tham như số tỉ lệ trộn, số thế hệ để tìm được bộ tham số phù hợp.
- Huấn luyện mô hình Wav2vec2.0 trên khung huấn luyện cải tiến. 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) trên benchmark VIVOS và Common Voice Vi.

KẾT QUẢ MONG ĐỢI

- 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 benchmark VIVOS và dưới 10% trên benchmark 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: labeling unsegmented sequence data with recurrent neural networks. *ICML 2006*: 369-376