



THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo:
<https://www.youtube.com/watch?v=y05-qBDez1k>
- Link slides:
https://github.com/quangtuan-0504/CS519_PPNCCKH/blob/main/slides.pdf

<ul style="list-style-type: none">● Họ và Tên: Hồ Thanh Tịnh● MSSV: 20520813 	<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: ??● Số câu hỏi QT của cả nhóm: ??● Link Github: https://github.com/quangtuan-0504/CS519_PPNCCKH.git● Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:<ul style="list-style-type: none">○ Chỉnh sửa phần trình bày, nội dung báo cáo, Slide và Poster○ Làm video YouTube
<ul style="list-style-type: none">● Họ và Tên: Nguyễn Đăng Quang Tuấn● MSSV: 20522116 	<ul style="list-style-type: none">● Lớp: CS519.O11● Tự đánh giá (điểm tổng kết môn): 9/10● Số buổi vắng: 200● Số câu hỏi QT cá nhân: 10● Số câu hỏi QT của cả nhóm: 15● Link Github: https://github.com/quangtuan-0504/CS519_PPNCCKH.git● Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:<ul style="list-style-type: none">○ Lên ý tưởng cho đồ án○ Tìm hiểu nội dung kiến thức.○ Viết báo cáo đồ án, viết slide và poster○ Làm video YouTube

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

XÂY DỰNG MÔ HÌNH NHẬN DẠNG GIỌNG NÓI TIẾNG VIỆT HIỆU QUẢ VỚI WAV2VEC 2.0

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

SPEECH RECOGNITION MODEL FOR VIETNAMESE USING WAV2VEC 2.0

TÓM TẮT (Tối đa 400 từ)

Bài toán nhận dạng giọng nói là một bài toán phổ biến và quan trọng, nó được sử dụng để xây dựng nhiều hệ thống hữu ích. Đã có rất nhiều phương pháp giải quyết tốt bài toán này, nhưng chúng thường có nhược điểm là cần rất nhiều dữ liệu gán nhãn để huấn luyện và việc dán nhãn rất tốn kém. Wav2vec 2.0 là một mô hình ra đời để giải quyết điều đó bằng cách sử dụng self-supervised learning, một mô hình chỉ cần một lượng nhỏ dữ liệu đã gán nhãn. Với lượng dữ liệu gán nhãn được đưa vào huấn luyện chỉ bằng 1/100 các mô hình thông thường, độ chính xác của nó có thể đánh bại những model hiện đại nhất. Nhưng các pretrained model Wav2vec2 hiện tại đa số dành cho Tiếng Anh, chính vì thế, trong nghiên cứu này chúng tôi tinh chỉnh Wav2vec2 trên bộ dữ liệu Tiếng Việt BKAI 2023 VN-SLU, bộ dữ liệu với chủ đề “Hiểu ngôn ngữ giọng nói tiếng Việt”, sau đó thực hiện đánh giá mô hình với thang đo Word Error Rate.

GIỚI THIỆU

Mạng neural thường có hiệu suất tốt khi có một lượng lớn dữ liệu đào tạo được gán nhãn. Tuy nhiên, trong nhiều trường hợp, việc thu thập dữ liệu được gán nhãn còn rất khó khăn. Do đó, self supervised learning đã xuất hiện như một giải pháp tận dụng lượng lớn dữ liệu không gán nhãn, bằng cách học các đặc trưng tổng quát từ dữ liệu không nhãn đó, rồi từ đó điều chỉnh mô hình trên dữ liệu có nhãn. Điều này đã đạt được thành công đặc biệt trong xử lý ngôn ngữ tự nhiên [1] và là một lĩnh vực nghiên cứu tích cực trong thị giác máy tính [2].

Trong đề tài này giới thiệu mô hình Wav2vec 2.0, một framework cho mô hình Self Supervised Learning học được các đặc trưng (representations) từ dữ liệu âm thanh thô để giải quyết cho bài toán speech2text. Phương pháp của đề tài là mã hóa âm thanh giọng nói thông qua một mạng CNN đa lớp thành các đặc trưng.

Các đặc trưng này được đưa vào một mạng Transformer để xây dựng thành các đặc trưng có ngữ cảnh. Mô hình sau đó được huấn luyện bởi Contrastive Task [5], trong đó, các đặc trưng có nhãn sẽ được phân biệt với các đặc trưng nhiễu khác.

Một phần quan trọng của quá trình huấn luyện, là việc học các đặc trưng đã được rời rạc hoá [6] thông qua một hàm phân phối Gumbel Softmax [7]. Điều này giúp biểu diễn các đặc trưng ẩn trong contrastive Task một cách hiệu quả, vượt trội hơn so với những đặc trưng không được rời rạc hoá trong các kỹ thuật truyền thống. Sau khi pre-training trên dữ liệu không nhãn, mô hình sẽ tiếp tục được fine-tune trên dữ liệu có nhãn sử dụng CTC Loss [8].

Wav2vec 2.0 có khả năng nhận dạng giọng nói hiệu quả chỉ với lượng dữ liệu có nhãn cực thấp, đạt tỷ lệ lỗi 4.8/8.2 WER chỉ với 10 phút dữ liệu có nhãn từ bộ Librispeech. Mô hình này cũng thiết lập kỷ lục mới trong nhận dạng âm vị, vượt trội so với các phương pháp trước đó [11].

Để kiểm chứng hiệu quả của mô hình này trên ngôn ngữ Tiếng Việt, chúng tôi sử dụng pretrained model đã được huấn luyện trên dữ liệu Tiếng Việt không gán nhãn và tinh chỉnh nó trên bộ dữ liệu Tiếng Việt khá nhỏ đã được gán nhãn, sau đó đánh giá lại mô hình bằng chính thang đo WER[12].

Nhóm đã xây dựng một hệ thống thử nghiệm để đánh giá độ hiệu quả của mô hình. Một hệ thống quản lý nhà thông minh với khả năng điều khiển các thiết bị như đèn, cửa, và các thiết bị khác thông qua giọng nói. trong việc kiểm soát nhà thông minh thông qua một hệ thống IoT sử dụng Arduino.

- Input: Đoạn audio, âm thanh có tiếng nói tiếng Việt ở dạng .mp3, .wav,..., được thu âm trực tiếp từ micro.
- Output: Văn bản chứa nội dung đoạn âm thanh tương ứng.

MỤC TIÊU

1. Tìm hiểu các kỹ thuật trong mô hình Wave2vec 2.0 tạo nên hiệu suất vượt trội của mô hình.
2. Tinh chỉnh mô hình trên bộ dữ liệu Tiếng Việt và đánh giá.
3. Xây dựng một hệ thống thử nghiệm nhỏ để cho thấy tính thực tiễn của bài toán ASR.

NỘI DUNG VÀ PHƯƠNG PHÁP

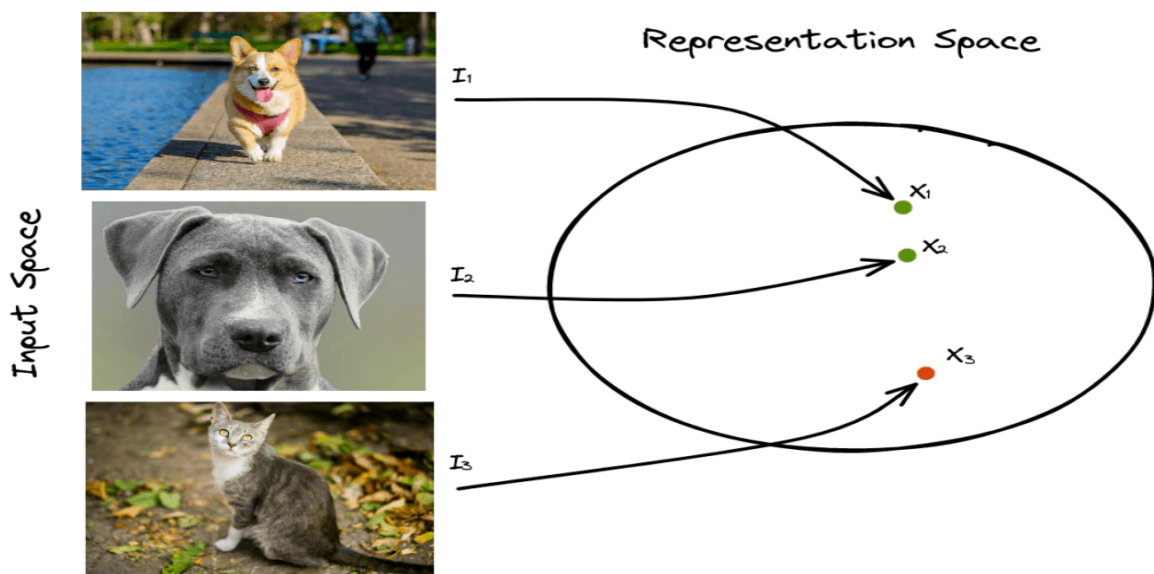
- Để giải quyết bài toán nhận dạng giọng nói với chỉ ít dữ liệu gán nhãn. Mô hình phải học được thông tin từ dữ liệu không gán nhãn, đây là khả năng đồng thời là ưu điểm chính của Self Supervised Learning cụ thể hơn là Contrastive Learning.

Mô hình Wav2Vec 2.0 được huấn luyện theo hai giai đoạn, giai đoạn đầu tiên mô hình sử dụng Feature Encoder giảm kích thước của dữ liệu không gán nhãn và tạo ra các vector đặc trưng. Sau đó Contrastive Learning biến đổi các vector đặc trưng này theo hai cách khác nhau:

- Context Network sử dụng Transformer Encoder để xử lý vector đặc trưng và thêm thông tin vị trí vào đầu vào.

- Quantization Module giúp chuyển đổi các giá trị liên tục thành một tập hữu hạn các giá trị rời rạc, sử dụng Gumbel-softmax. Đồng thời, trong phép biến đổi này cũng sử dụng Diversity Loss để đảm bảo sự đồng đều trong việc sử dụng các token từ codebook với khả năng như nhau.

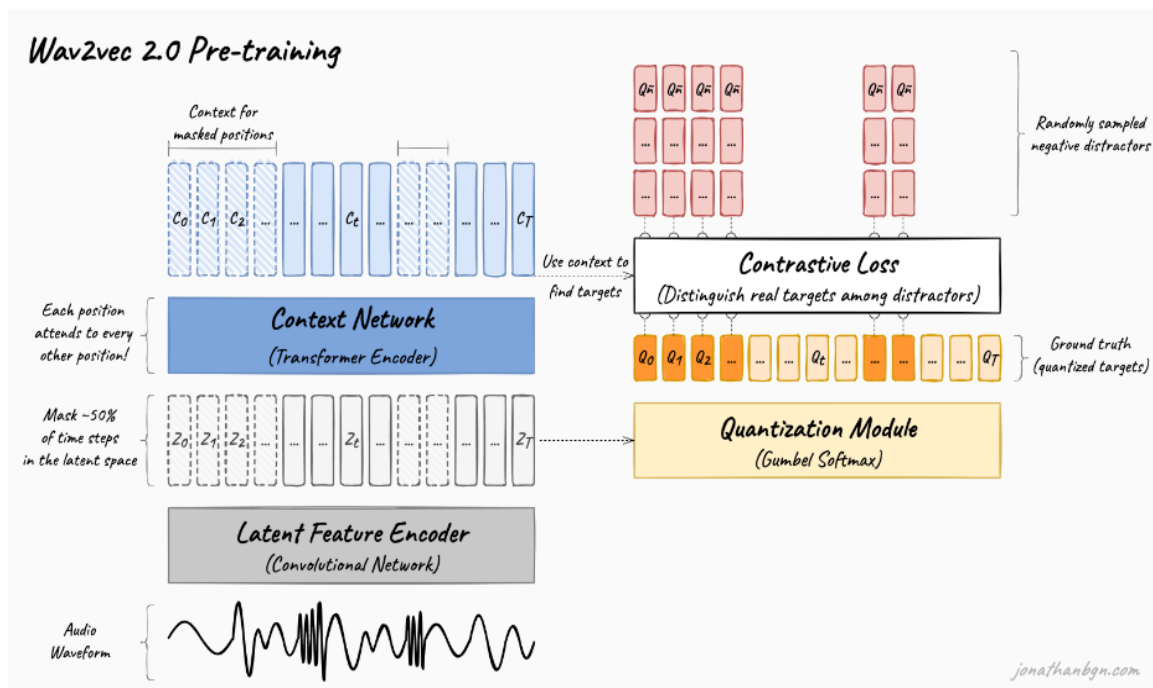
Sau đó mô hình nhận biết liệu hai biến đổi có thuộc cùng một đối tượng hay không bằng cách sử dụng Contrastive Loss. Để đảm bảo Context Vector chọn đúng Quantized Vector tương ứng.



Hình 1. Minh họa cho contrastive learning

- Mô hình còn sử dụng masking để che đi một số vector đặc trưng để tăng khả năng dự đoán token tiếp theo, giúp mô hình học được biểu diễn chất lượng cao trong các điểm quan trọng của âm thanh.

- Những kỹ thuật này mang lại hiệu suất vượt trội cho mô hình, đặc biệt là khi xử lý giọng nói với khả năng biểu diễn rời rạc dữ liệu âm thanh.



Hình 2: Kiến trúc mô hình Wav2vec 2.0

Trong giai đoạn thứ hai, mô hình sau khi được huấn luyện với dữ liệu không gán nhãn sau giai đoạn đầu tiên tiếp tục được huấn luyện bằng dữ liệu được gán nhãn thông qua một cấu trúc tương tự. Tuy nhiên, ở cấu trúc này, quantization module không được sử dụng. Thay vào đó, một lớp fully connected được khởi tạo ngẫu nhiên được thêm vào phía sau context network. Sau đó, mô hình này được tinh chỉnh với hàm mất mát CTC loss [8].

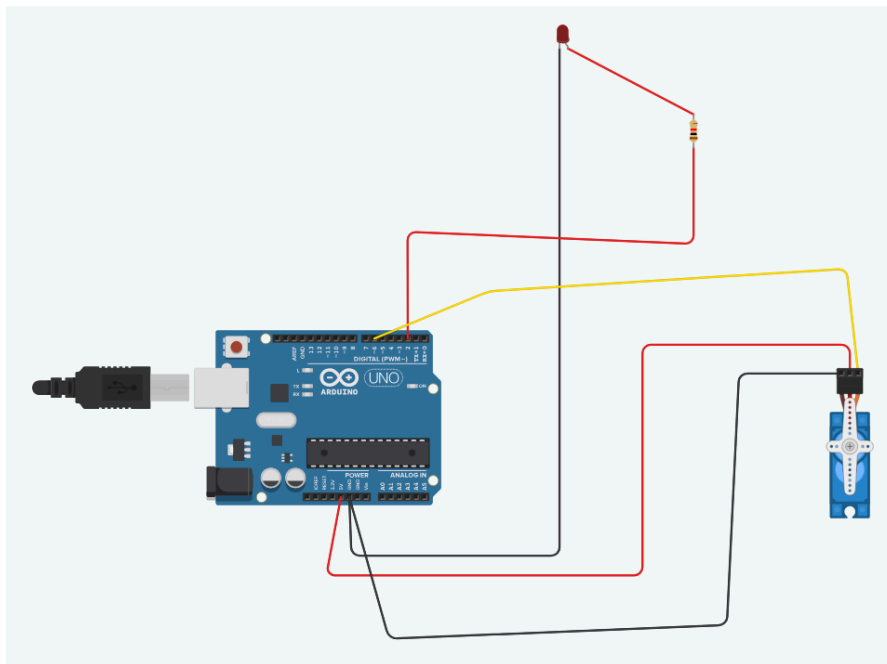
Để kiểm tra hiệu suất trên bộ dữ liệu Tiếng Việt một cách rõ ràng thông qua hai giai đoạn của mô hình:

- Ở giai đoạn 1: Sử dụng mô hình pretrained nguyenvulebinh/wav2vec2-base-vietnamese-250h được lấy từ Hugging Face, mô hình này đã được pre-trained trên 13k giờ giọng nói tiếng Việt được lấy từ youtube không gán nhãn.
- Ở giai đoạn 2: Thực hiện tinh chỉnh mô hình pretrained này với hai bộ dữ liệu Tiếng Việt gán nhãn. Bộ dữ liệu đầu tiên gồm 250 giờ data được gán nhãn của bộ dữ liệu VLSP ASR. Và bộ dữ liệu thứ hai gồm 7490 mẫu âm thanh có thời lượng từ 3 đến 5 giây dữ liệu gán nhãn được lấy từ cuộc thi BKAI2023 với chủ đề “hiểu ngôn ngữ Tiếng Việt trong nhà thông minh”. Tất cả đều có sample rate là 16khz.

Việc làm này sẽ kiểm nghiệm xem Wav2vec 2.0 có thực sự tốt với lượng nhỏ dữ liệu gán nhãn như bài báo đã nói không và để biết Wav2vec 2.0 hoạt động như thế nào đối với Tiếng Việt

Cuối cùng, nhằm cho thấy tính ứng dụng của bài toán ASR và là bàn đạp để xây dựng các hệ thống thực tiễn, chúng tôi xây dựng một hệ thống điều khiển các thiết bị bằng giọng nói với sự hỗ trợ của Arduino.

Để chuẩn hóa văn bản thành các lệnh cụ thể, giúp hệ thống tự động hiểu ngữ cảnh và thực hiện tương tác hiệu quả hơn, nhóm đã sử dụng bổ sung mô hình ngôn ngữ để chuyển đổi văn bản đầu vào như "bật giúp cho tôi đèn và đóng luôn cái cửa giùm tôi" thành lệnh "bật đèn, đóng cửa," tăng khả năng tự động hóa của hệ thống.



KẾT QUẢ MONG ĐỢI

- Hiểu được các kỹ 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.

TÀI LIỆU THAM KHẢO

- [1] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In Proc. of ACL, 2018.
- [2] P. Bachman, R. D. Hjelm, and W. Buchwalter. Learning representations by maximizing mutual information across views. In Proc. of NeurIPS, 2019.
- [3] D. Jiang, X. Lei, W. Li, N. Luo, Y. Hu, W. Zou, and X. Li. Improving

transformer-based speech recognition using unsupervised pre-training. arXiv, abs/1910.09932, 2019.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv, abs/1810.04805, 2018.

[5] A. van den Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. arXiv, abs/1807.03748, 2018.

[6] D. Harwath, W.-N. Hsu, and J. Glass. Learning hierarchical discrete linguistic units from visually-grounded speech. In Proc. of ICLR, 2020.

[7] A. Baevski, S. Schneider, and M. Auli. vq-wav2vec: Self-supervised learning of discrete speech representations. In Proc. of ICLR, 2020.

[8] A. Graves, S. Fernández, and F. Gomez. Connectionist temporal classification: Labeling unsegmented sequence data with recurrent neural networks. In Proc. of ICML, 2006.

[9] D. Jiang, X. Lei, W. Li, N. Luo, Y. Hu, W. Zou, and X. Li. Improving transformer-based speech recognition using unsupervised pre-training. arXiv, abs/1910.09932, 2019.

[10] A. Baevski, M. Auli, and A. Mohamed. Effectiveness of self-supervised pre-training for speech recognition. arXiv, abs/1911.03912, 2019.

[11] D. S. Park, Y. Zhang, Y. Jia, W. Han, C.-C. Chiu, B. Li, Y. Wu, and Q. V. Le. Improved noisy student training for automatic speech recognition. arXiv, abs/2005.09629, 2020.

[12] WER Wikipedia

[13] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, Michael Auli. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. arXiv:2006.11477v3 [cs.CL] 22 Oct 2020.