BÁO CÁO ĐỒ ÁN CUỐI KỲ

Môn học: CS519 - PHƯƠNG PHÁP LUẬN NCKH

Lớp: CS519.O11

GV: PGS.TS. Lê Đình Duy

Trường ĐH Công Nghệ Thông Tin, ĐHQG-HCM

NHẬN DẠNG GIỌNG NÓI TIẾNG VIỆT TỰ ĐỘNG

Nguyễn Đặng Quang Tuấn - 20522116
Hồ Thanh Tịnh - 20520813

Thông tin nhóm

- Link Github của nhóm: https://github.com/quangtuan-0504/CS519_PPNCKH.git
- Link YouTube video: https://www.youtube.com/watch?v=y05-qBDez1

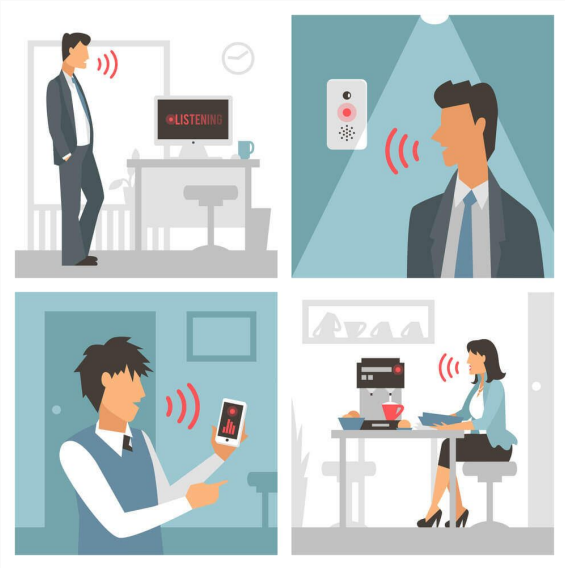
Nguyễn Đặng Quang Tuấn

Hồ Thanh Tịnh

Tóm tắt

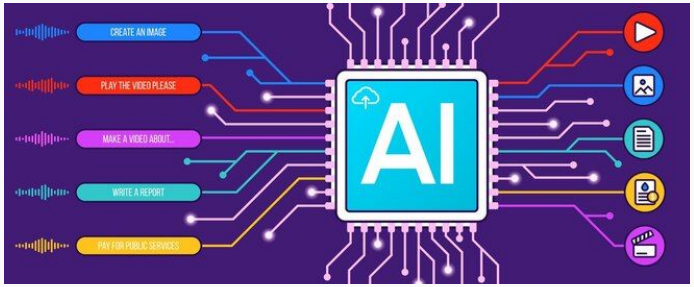
INPUT: Đoạn âm thanh có thể chứa giọng nói.

OUTPUT: Văn bản của đoạn âm thanh giọng nói Tiếng Việt.



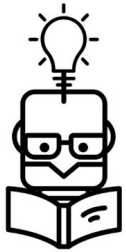
Giới thiệu

Các mô hình supervised- learning cần lượng lớn labeled data



Tồn tại nhiều nguồn dữ liệu giọng nói unlabeled data

- Làm thế nào để tận dụng lượng unlabeled data?



Wave2vec 2.0: Self Supervised Learning model

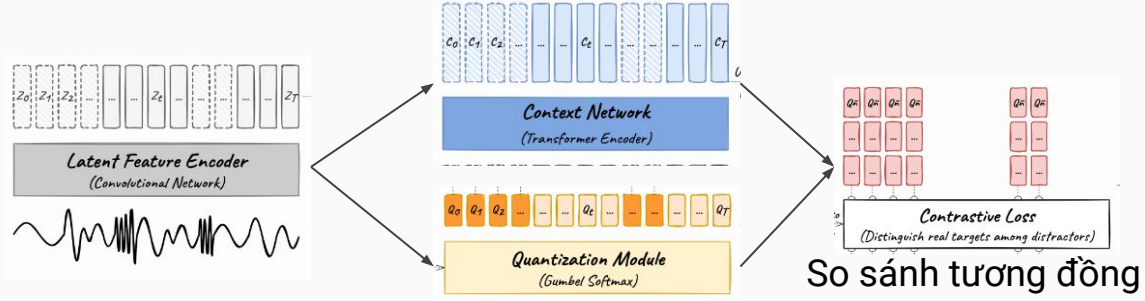
- Độ chính xác đánh bại các mô hình SOTA chỉ với lượng nhỏ dữ liệu gán nhãn.
- Nhưng Wav2vec 2.0 có hoạt động tốt với Tiếng Việt hay không?

Mục tiêu

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

Nội dung và phương pháp

- Contrastive Learning: Giai đoạn 1, huấn luyện trên dữ liệu không có nhãn

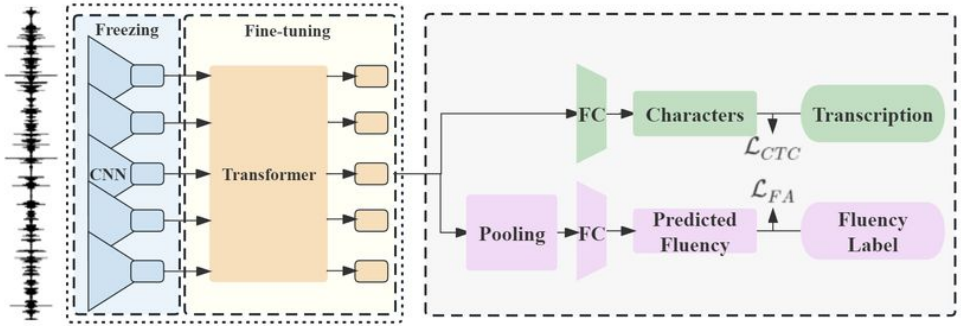


Nén và tạo đặc trưng

Biến đổi đặc trưng

So sánh tương đồng

- Contrastive Learning: Giai đoạn 2, fine tune trên dữ liệu có nhãn



Nội dung và phương pháp

Thử nghiệm 2 giai đoạn trên tập dữ liệu Tiếng Việt:

- Pre-train giai đoạn 1 trên 13k giờ dữ liệu không nhãn trích từ youtube

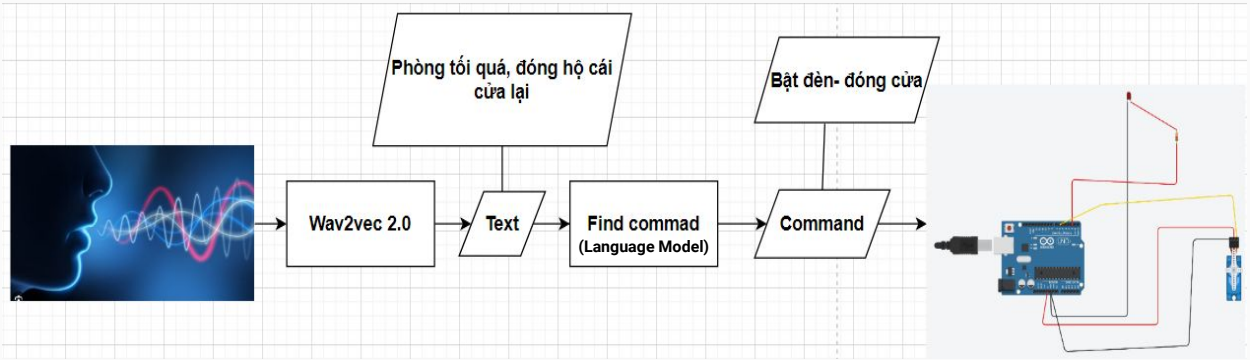
- Fine tune giai đoạn 2 trên 250 giờ dữ liệu có nhãn VLSP

VLSP

- Fine tune với 7490 mẫu âm thanh đã gán nhãn thời lượng từ 2-5s, từ BKAI2023-SLU.

BKAI

- Hệ thống thử nghiệm



Kết quả mong đợi

- Mô hình huấn luyện đầy đủ 2 giai đoạn trên tập dữ liệu Tiếng Việt.
- Mô hình được tinh chỉnh 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] A. Baevski, S. Schneider, and M. Auli. vq-wav2vec: Self-supervised learning of discrete speech representations. In Proc. of ICLR, 2020.

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

[5] 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.

[6] 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.

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