# Báo cáo
# Mô hình Wave2Vec 2.0

1. **Nguyên nhân:**

- Trong quá trình kiểm thử và thực hiện trên dữ liệu thực, mô hình nhận diện giọng nói còn gặp nhiều sai sót.
- Nguyên nhân có thể xuất phát do mô hình còn đơn giản.
- Ngoài ra có thể do tập dữ liệu là dữ liệu tiếng Anh không phải tiếng Việt.

2. **Đề xuất giải pháp:**

- Giải pháp 1: Sử dụng mô hình mới -> Sử dụng mô hình Wave2Vec 2.0 .
- Giải pháp 2: Sử dụng tập dữ liệu mới -> Tập dữ liệu VIVOS chứa dữ liệu các câu nói tiếng Việt.

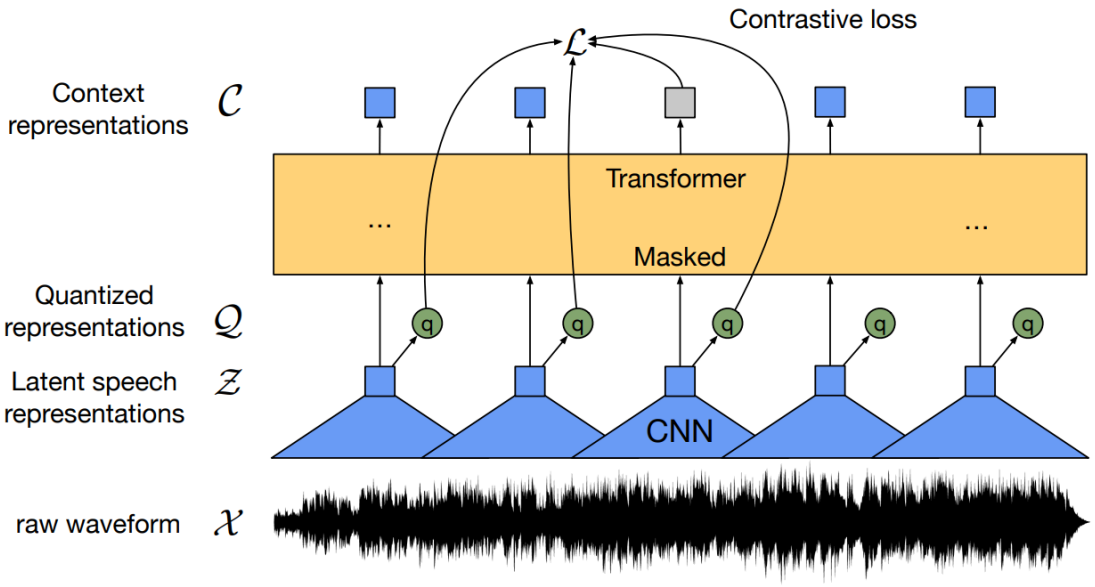
⇒ **Áp dụng cả 2 giải pháp.**

3. **Tập dữ liệu VIVOS:**

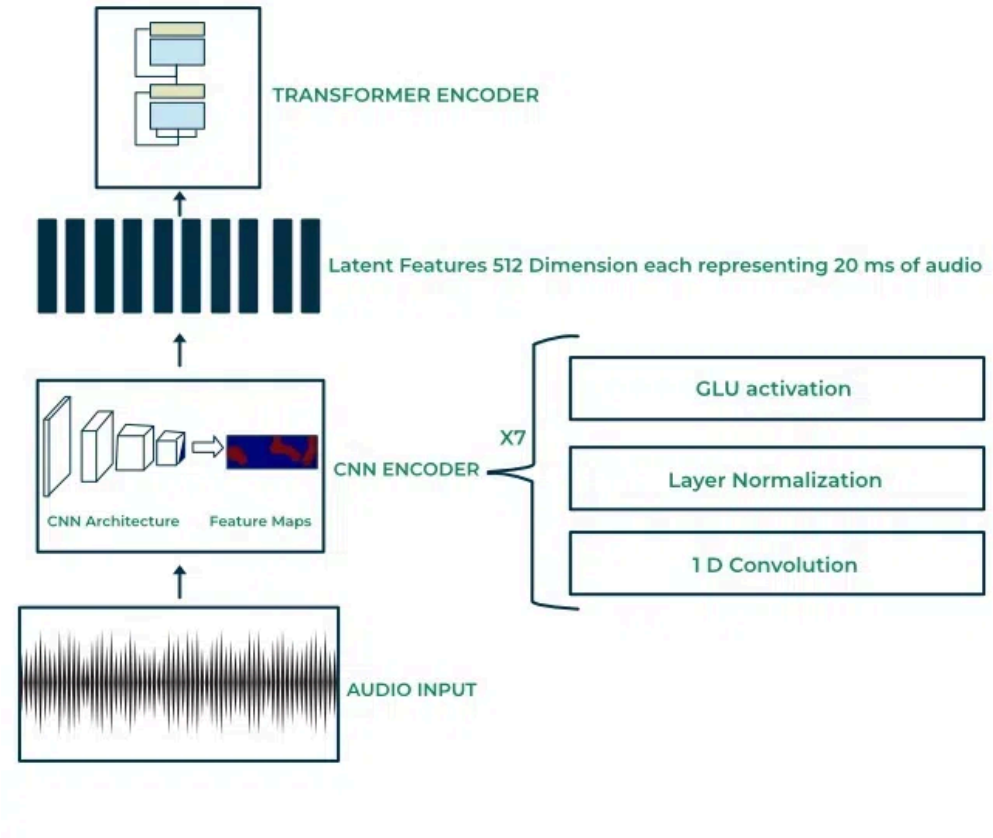
- Nguồn: HuggingFace
- Link: [vivos · Datasets at Hugging Face]
- Dữ liệu bao gồm speaker_id, path, sentence.
  + speaker_id: Mã người nói
  + path: Đường dẫn tới audio file
  + sentence: Câu nói trong video
- Để áp dụng cho bài toán Voice Recognition thì chỉ sử dụng speaker_id và path.

4. **Mô hình Wave2Vec 2.0:**

- Mô hình Wave2Vec 2.0 (Wave2Vec2) là mô hình xử lý âm thanh phổ biến nhất hiện nay.
- Mô hình pre-train: *facebook/wav2vec2-base-960h*
- Đầu vào của mô hình là âm thanh thô.
- Đầu ra là một embedding vector với số chiều là (786)
- Mô hình: Âm thanh thô đi vào sẽ đi qua 1 lớp CNN Encoder để xuất ra 1 mảng feature với số chiều là 512 với mỗi đoạn âm thành 20ms. Cuối cùng đi qua lớp transformer để đưa ra embedding vector.
- Mô hình:




Hình1. Mô hình Wave2Vec 2.0 (1)




Hình2. Mô hình Wave2Vec 2.0 (2)

a. **Áp dụng mô hình Wave2Vec 2.0 vào bài toán Voice Recognition:**

- Cách thức hoạt động: Lấy ra được embedding vector từ mô hình Wave2Vec 2.0, tiếp đến sử dụng Kmean để phân loại.
- Nhận xét:
  + Kết quả thu được ban đầu khá khả quan với kết kết quả thu được trên tập test của mỗi người đều trên 80% nhưng sau khi tiếp tục huấn luyện kết quả lại giảm đi đáng kể chỉ còn 37,5%.
  + Nguyên do dẫn đến giảm có thể do việc huấn luyện quá lâu dẫn đến không thể hoàn thành huấn luyện mô hình trọn vẹn 1 epoch mỗi lần huấn luyện nên em đã lặp đi lặp lại các lần huấn luyện và lưu mô hình mỗi lần hoàn thành 1 batch (mỗi batch là 1 mẫu) dẫn đến mô hình quá khớp. Ngoài ra có thể do sử dụng margin cho TripletLoss khá nhỏ (chỉ 0.1).
  + Giải pháp kết tiếp: Sẽ huấn luyện lại mô hình và lưu mô hình mỗi lần hoàn thành 1 epoch.

b. **Định hướng áp dụng mô hình Wave2Vec 2.0 vào bài toán Voice DeepFake Detection:**

- Cách thức hoạt động: Xây dựng mô hình Wave2Vec2ClassificationModel với đầu vào là âm thanh thô. Sử dụng mô hình Wave2Vec 2.0 để đưa ra embedding vector, sau đó sử dụng 1 lớp Full Connection để đưa về vector 2 chiều chứa xác suất giọng thật và xác xuất giả.
- Hiện tại mô hình đang ở ý tưởng và chưa được huấn luyện.

Nguồn tham khảo:


- Speech Recognition | Papers With Code
- 2006.11477.pdf (arxiv.org)
- facebook/wav2vec2-base-960h · Hugging Face
- vivos · Datasets at Hugging Face
- Audio classification (huggingface.co)