BÁO CÁO

Speech Recognization using Whisper and PhoWhisper

1. **Bài toán nhận dạng giọng nói:**

- Nhận dạng giọng nói, còn được gọi là nhận dạng giọng nói tự động (ASR) hay chuyển giọng nói thành văn bản (STT) là công nghệ cho phép máy tính nhận dạng và chuyển đổi ngôn ngữ nói thành văn bản.

- Công nghệ nhận dạng giọng nói sử dụng AI và các mô hình học máy để xác định và phiên âm chính xác các giọng, phương ngữ và mẫu giọng nói khác nhau.

- Các ứng dụng của nhận dạng giọng nói có thể kể đến như ghi chép nội dụng trong một cuộc họp, tổng đài hỗ trợ tự động, … .

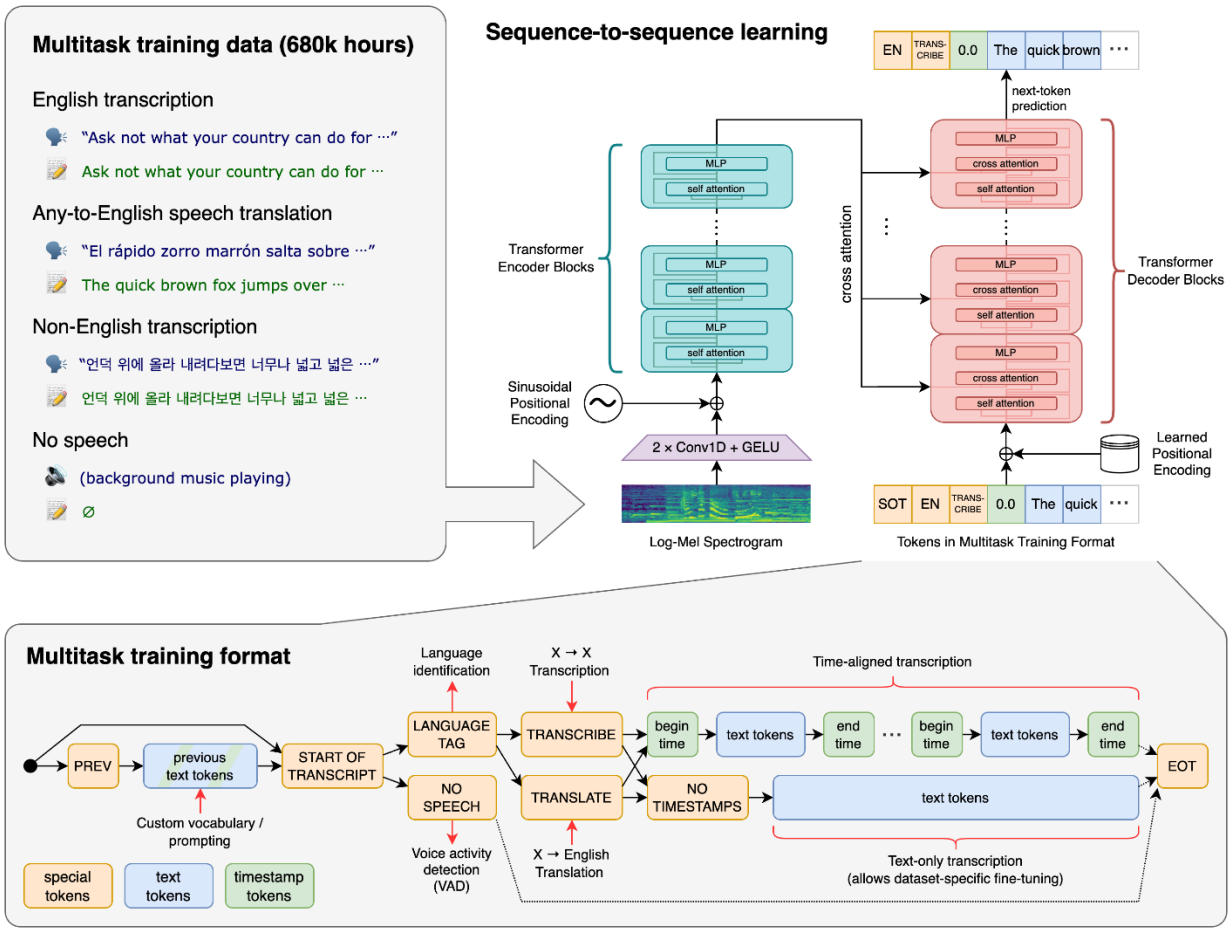
2. **Mô hình Whisper:**

Model	Layers	Width	Heads	Parameters
Tiny	4	384	6	39M
Base	6	512	8	74M
Small	12	768	12	244M
Medium	24	1024	16	769M
Large	32	1280	20	1550M

a. Tiền xử lý:

- Mô hình Whisper khi đọc 1 file audio sẽ chuyển audio đó thành log-Mel spectrogram với n_mels = 80.
- Trong đó:
    - n_mels là số lượng Mel spectrogram được trích xuất ra từ file âm thanh.

b. Mô hình:



- Mô hình Whisper của OpenAI được cấu trúc dựa trên kiến trúc máy biến áp bộ mã hóa-giải mã. Cụ thể, mô hình này bao gồm:
    - Bộ mã hóa (Encoder/Transformer Encoder): Bộ mã hóa chịu trách nhiệm chuyển đổi dữ liệu âm thanh đầu vào thành biểu diễn vector. Dữ liệu âm thanh đầu vào được chia thành các đoạn dài 30 giây, được chuyển đổi thành biểu đồ quang phổ log-Mel, và sau đó được chuyển đến bộ mã hóa.
    - Bộ giải mã (Decoder/Transformer Decoder): Bộ giải mã nhận biểu diễn vector từ bộ mã hóa và chuyển đổi chúng thành văn bản. Bộ giải mã có khả năng tạo ra các bản ghi với khả năng đọc nâng cao và dấu thời gian ở cấp độ cụm từ.
- Cả hai thành phần này làm việc cùng nhau để chuyển đổi âm thanh thành văn bản trong mô hình Whisper. Kiến trúc end-to-end này cho phép mô hình học cách ánh xạ trực tiếp từ dữ liệu âm thanh đầu vào đến văn bản đầu ra mà không cần bất kỳ giai đoạn tiền xử lý hoặc hậu xử lý nào.

3. Mô hình PhoWhisper:

- Mô hình PhoWhisper được fined-tuning trên mô hình đa ngôn ngữ của mô hình Whisper và được phát triển bởi VinAI trên tập dữ liệu 844 giờ ghi âm đa dạng giọng nói tiếng Việt.
- Trương tự Whisper PhoWhisper cũng có 5 loại model: Tiny, Base, Small, Medium, Large

4. Kết quả thu được:

	model	detection	insertion	total_time
0	tiny whisper	0.762945	0.335356	121.304846
1	base whisper	0.860437	0.221278	39.251880
2	small whisper	0.883900	0.173948	54.972825
3	medium whisper	0.909790	0.146036	90.673369
4	large whisper	0.911408	0.149272	123.028449
5	largev2 whisper	0.904126	0.171521	125.639401
6	largev3 whisper	0.911408	0.149272	123.609138

Kết quả của các mô hình whisper.

	model	detection	insertion	total_time
0	tiny phowhisper	0.947006	0.124191	33.487694
1	base phowhisper	0.948220	0.115291	25.329878
2	small phowhisper	0.968447	0.097492	41.617583
3	medium phowhisper	0.965615	0.095874	86.714818
4	large phowhisper	0.977346	0.091424	140.098449

Kết quả của các mô hình phowhisper.

- Chú thích:

- model: tên mô hình được sử dụng.
- correctly_determined: Số từ phát hiện đúng trên tổng số từ được nói.
- wrong_determined: Số từ phát hiện đúng trên tổng số từ được nói.
- total_time: tổng thời lượng để chuyển tất cả ghi âm thành chữ (Sử dụng T4 GPU của google colab).

5. Nhận xét:

- Dữ liệu kiểm thử:
  - Mức độ dễ nhận diện nhất đến khó nhận diện nhất: Trí, Phát, Đạt, Tuấn.
  - Dữ liệu Trí, Phát, Đạt, Tuấn là các câu hoàn chỉnh, âm thu được khá rõ ràng.
  - Dữ liệu Tuấn_prev khó nhận diện nhất do âm thanh hơi nhỏ.
- Whisper:
  - Mô hình Tiny của Whisper đa phần nhận diện từ bị sai và sai ngôn ngữ.
  - Mô hình Base của Whisper nhận diện khá tốt ở dữ liệu Trí, nhận diện sai nhiều ở các tập kiểm thử khác, gặp sai các dấu âm.
  - Mô hình Small của Whisper nhận diện khá tốt ở dữ liệu Trí và Phát, nhưng nhận diện sai nhiều ở dữ liệu Đạt và Tuấn, nhận diện sai chữ "gara" thành "gà ra".
  - Mô hình Medium nhận diện tốt ở dữ liệu Trí và Phát, nhận diện khá tốt ở tập dữ liệu Đạt nhưng nhận diện sai nhiều ở tập dữ liệu Tuấn.
  - Mô hình Large nhận diện tốt ở dữ liệu Trí, Phát và Đạt nhưng lại nhận diện sai ở tập dữ liệu Tuấn.
- PhoWhisper:
  - Mô hình Tiny của PhoWhisper nhận diện tốt ở dữ liệu Trí và Phát, nhận diện khá tốt ở tập dữ liệu Đạt nhưng nhận diện sai ở tập dữ liệu Tuấn.
  - Mô hình Base của PhoWhisper nhận diện tốt ở dữ liệu Trí và Phát, nhận diện khá tốt ở tập dữ liệu Đạt nhưng nhận diện sai nhiều ở tập dữ liệu Tuấn.
  - Mô hình Small của PhoWhisper nhận diện tốt ở dữ liệu Trí và Phát, nhận diện khá tốt ở tập dữ liệu Đạt nhưng nhận diện sai nhiều ở tập dữ liệu Tuấn (có cải thiện hơn mô hình Base của PhoWhisper).
  - Mô hình Medium của PhoWhisper nhận diện tốt ở dữ liệu Trí và Phát, nhận diện khá tốt ở tập dữ liệu Đạt nhưng nhận diện tốt (có một số từ bị nhận diện sai như gara, đèn,...) ở tập dữ liệu Tuấn.
  - Mô hình Largecủa PhoWhisper nhận diện tốt ở dữ liệu Trí và Phát, nhận diện khá tốt ở tập dữ liệu Đạt nhưng nhận diện tốt (có một số từ bị nhận diện sai như gara, đèn,...) ở tập dữ liệu Tuấn (Tương tự Mô hình medium).