HUST

ĐẠI HỌC BÁCH KHOA HÀ NỘI

HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

ONE LOVE. ONE FUTURE.

ĐẠI HỌC
BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY
OF SCIENCE AND TECHNOLOGY

SPEECH TECHNOLOGY

Automatic Speech Recognition Challenge
Group 12

ONE LOVE. ONE FUTURE.

Giới thiệu vấn đề

Giới thiệu vấn đề

Nhận dạng tiếng nói tự động (ASR) là công nghệ chuyển đổi âm thanh tiếng nói thành văn bản.

Ứng dụng: trợ lý ảo, chuyển đổi giọng nói thành văn bản, hỗ trợ người khuyết tật, nhập liệu nhanh,...

$\Rightarrow$Mục tiêu: Xây dựng hệ thống nhận dạng tiếng nói tiếng Việt với độ chính xác cao (WER thấp).

Tổng quan pipeline

Tổng quan pipeline

Các bước chính của pipeline:

- Cài đặt môi trường và thư viện
- Tiền xử lý dữ liệu audio và transcript
- Chuẩn bị đặc trưng và tokenizer
- Xây dựng, huấn luyện và đánh giá mô hình

Môi trường và thư viện

Cài đặt môi trường

- transformers: Cho mô hình Whisper
- datasets[audio]: Xử lý dữ liệu âm thanh
- torchaudio: Xử lý âm thanh
- deepfilternet: Lọc nhiễu
- evaluate, jiwer: Đánh giá mô hình
- bitsandbytes: Tối ưu bộ nhớ
- PEFT: Cho fine-tuning hiệu quả

Tiền xử lý dữ liệu

DATASET

- Vietnamese Language and Speech Processing 2020 Speech Recognition Dataset
  - Kích thước: Khoảng 100 giờ ghi âm tiếng Việt
  - Định dạng: Cặp file audio (.wav) và transcript (.txt) tương ứng

Tiền xử lý

- Audio:
  - Đọc file, chuẩn hóa tần số lấy mẫu (16kHz)
  - Lọc nhiễu bằng DeepFilterNet
- Transcript:
  - Đọc file, chuẩn hóa text, loại bỏ ký tự thừa
- Tạo tập train/val/test:
  - 7000 mẫu train, 2000 mẫu validation, 1000 mẫu test

Chuẩn bị đặc trưng và tokenizer

- Sử dụng WhisperFeatureExtractor để trích xuất log-Mel spectrogram từ audio.
- Sử dụng WhisperTokenizer để mã hóa transcript thành label ids.
- Map dữ liệu sang định dạng phù hợp cho model.

Xây dựng và huấn luyện mô hình

Xây dựng và huấn luyện mô hình

- Sử dụng mô hình Whisper (openai/whisper-small), fine-tune cho tiếng Việt.
- Định nghĩa data collator, hàm tính WER, thiết lập Trainer.

```
# Khởi tạo base model
model = WhisperForConditionalGeneration.from_pretrained("whisper-small-vi")

# Cấu hình generation
model.config.forced_decoder_ids = None
model.config.suppress_tokens = []

model.generation_config.language = "vietnamese"
model.generation_config.task = "transcribe"
```

Kết quả

Kết quả

Kết quả sau quá trình train

- Training Loss và Validation Loss đều giảm qua các epoch, cho thấy mô hình học tốt và không bị overfit.
- Word Error Rate (WER) giảm từ 21.99% xuống 20.02% sau 2 epoch, chứng tỏ mô hình cải thiện khả năng nhận dạng tiếng nói.
- Kết quả WER ~20% là khá tốt cho bài toán ASR tiếng Việt với dữ liệu và cấu hình hiện tại.

Epoch	Training Loss	Validation Loss	Wer
1	0.303000	0.517939	21.987348
2	0.167400	0.481668	20.022630

HUST

hust.edu.vn fb.com/dhbkhn

THANK YOU !