Placement Project - Augnito AI

Jithu J

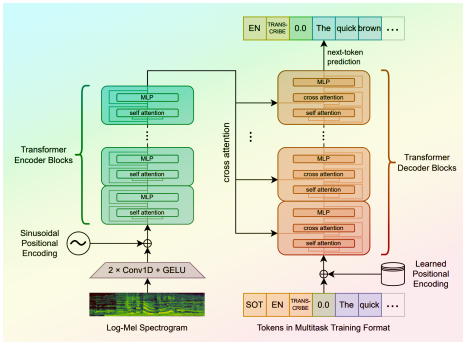
November 2024

1 Introduction

Whisper is an Automatic Speech Recognition (ASR) Model for recognition and transcription of speech. It was created by OpenAI and first released as open-source software in September 2022.
Key Features of Whisper Model include- multilingual support, Speech Recognition and Translation, Robustness to accents and background noise,

2 Literature

Whisper model is a supervised automatic speech recognition system developed by Open Ai.Whisper was trained on an extensive dataset of 680,000 hours of multilingual and multitask audio. Whisper uses a sequence-to-sequence Transformer architecture with an encoder-decoder structure, designed for flexibility across transcription, translation, and other speech-processing tasks. This setup enables the model to handle complex, multi-language audio input, including speech in noisy or otherwise challenging environments.
Audio is converted to a log-Mel spectrogram—a standard audio feature representation that encodes the frequency and intensity of sound over time.This spectrogram is divided into 25ms windows with a 10ms overlap, producing an 80-channel Mel-spectrogram. The input is normalized globally to improve the model's generalization and robustness.The model Architecture follows.



The dataset used for fine tuning is a kaggle dataset named medical-speech-transcription-and-intent. It contains nearly 6661 audio files split into 3 folders(Test, Train, Validate).It also consit of metadata as a csv file. It is audio of patients and its transcription. The following article acted as a reference for model training: https://huggingface.co/blog/fine-tune-whisper

3 Methodology

The various Whisper models and its features are given below:

Size	Parameters	English-only model	Multilingual model	Required VRAM	Relative speed
tiny	39 M	`tiny.en`	`tiny`	~1 GB	~10x
base	74 M	`base.en`	`base`	~1 GB	~7x
small	244 M	`small.en`	`small`	~2 GB	~4x
medium	769 M	`medium.en`	`medium`	~5 GB	~2x
large	1550 M	N/A	`large`	~10 GB	1x
turbo	809 M	N/A	`turbo`	~6 GB	~8x

With the help of the table the whisper-small model was chosen as a checkpoint for model training. The choice of whisper small is considering the training time and no of parameters. Since the task only requires english transcription the whisper model would work just fine.
First we loaded dataset from kaggle using kagglehub library. The dataset contained 3 folders and metadata in form of csv file. the test folder contained majority of the audio files 5800+ to be specific. hence only this data was used for training.
Data preprocessing:: The metadata was loaded to the notebook using dataset library from Huggingface platform. Load dataset() function was used. Then audio data was loaded similarly. We mapped audio data corresponding to the file name in metadata and dropped all the unnecessary coloumns to create the dataset keeping only audio and corresponding transcriptions. We used train test split to create training and testing splits(at test=0.2). The audio data is converted to mel spectrogram and the sentences are converted to tokens using whisperfeatureextracter and Whispertokenizer from transformers library respectively.
Importing checkpoints:: The whisper small model was imported from transformers library. The whisper model is trained on 680000 hours of language data. This model is used as a checkpoint. Further the task was supposed to be completed in colab notebook but since the notebook timeout happens for large model timing the model was finetuned locally. so inorder to move the model to gpu specific functions from pytorch library is used.
Model Training:: The model is trained on a nvidia RTX 4050 gpu. Training had 4000 steps across 14 epochs. The train split had 4716 audio files and test split had 1179 audio files. We use training loss, validation loss and word error

rate as evaluation metrices. The learning rate is set at 1e-5.

4 Result

The model took 6.30Hr to complete. The completed model was pushed to huggingface. It can be accessed at following directory:: https://huggingface.co/jithuj12344321/whisper-small-en. Best achieved word error rate is:: 3.9 on the given dataset.

Training result

Training Loss	Epoch	Step	Validation Loss	Wer
0.0165	3.3898	1000	0.0971	4.7860
0.0012	6.7797	2000	0.0905	4.1425
0.0001	10.1695	3000	0.0930	4.0138
0.0001	13.5593	4000	0.0931	3.9012

The model was loaded in a colab note book and tested on some audio files. The model transcribed the data without failure.
A sample audio recorded in real time of me saying the sentence "I have been experiencing severe pain in my lower back for the past two weeks. The pain started after I lifted a heavy object at work." was recorded and tested with the model. The model transcribed it impeccably. The result is shown in a separate notebook in the drive.