Voice AI - Automatic Speech Recognition

Introduction

Welcome to the Voice AI project, where our focus is on implementing a Speech-to-Text system for the Marathi language using the cutting-edge Hugging Face Whisper ASR models. This project is designed to deliver a robust Marathi speech-to-text transcription model with the ultimate goal of accurately transcribing spoken Marathi content.

Environment Setup

Ensure the following libraries are installed:

```
pip install --upgrade pip

pip install --upgrade git+https://github.com/huggingface/transformers.git accelerate datasets transformers soundfile librosa evaluate tensorboard

pip install pandas

pip install torch

pip install python-Levenshtein

pip install jiwer
```

Data Collection

This project relies on the Common Voice Marathi dataset from Mozilla Foundation for training and evaluating the Marathi Speech Recognition model. The dataset includes essential attributes such as a path column for easy access to audio files, an array column for numerical audio representations, sampling rate information, and a sentence column providing transcriptions for training the model on spoken Marathi sentences.

Data Preprocessing

- **Down-sampling Rate**: The original audio, recorded at 48000Hz, is down-sampled to 16000Hz. This step is essential for standardizing the audio data and reducing computational complexity.
- **Feature Extraction**: The Whisper models' feature extractor is employed to extract relevant features from the audio signals. These features play a crucial role in training the speech recognition model.
- **Tokenization**: It involves breaking down the audio data into smaller units (tokens). This process is fundamental for converting the continuous audio stream into a format suitable for machine learning algorithms.

Speech Recognition Models

- The primary architecture for this project is based on the Whisper ASR models, with a focus on both the Whisper Small and Large versions. These models are selected for their effectiveness in speech recognition tasks and will be used for comparison during the training and evaluation phases**.**
- Additionally, to enhance training efficiency and adapt the models to specific tasks, we apply PEFT (Probabilistic Exposure Fine-Tuning) and LORA (Layer-wise Adaptive Rate) adaptations. These techniques contribute to improving the overall performance and robustness of the Marathi Speech Recognition system.

Fine-Tuning

To optimize the performance of the models, fine-tuning strategies are applied. These strategies involve carefully adjusting model parameters and hyper parameters during the training process, ensuring the best possible outcomes for our specific speech recognition tasks.

Model Training

For model training, we leverage the powerful Seq2SeqTrainingArguments and Seq2SeqTrainer from the Hugging Face Transformers library. These components provide a robust framework for training our Speech Recognition models**.**

Step	Training Loss	Validation Loss	WER
25	19.471	35.393	100.00%
50	72.230	153.046	100.00%
100	84.854	110.924	100.00%

Model Evaluation

- The primary metrics for evaluating the performance are Word Error Rate (WER) and text similarity scores. The objective is to minimize WER, ensuring accurate transcription of Marathi speech.
- Before fine-tuning, we assessed the models on the provided test dataset. For the Whisper Large-v3 model, the calculated average WER was 0.74, with an average similarity score of 78.66.
- In comparison, the Whisper Small model showed an average WER of 3.38 and an average similarity score of 30.48. These metrics serve as benchmarks for gauging improvements during the fine-tuning process.

Challenges Faced

- **GPU and Storage Constraints**: Limited storage capacity in Google Colab posed a challenge, preventing additional fine-tuning steps due to insufficient space for model checkpoints and intermediate results. The free version of Google Colab provided inadequate GPU capacity, hindering the fine-tuning of larger and more complex models. This limitation impacted training efficiency and overall model performance.
- **Model Complexity vs Steps**: Balancing increased model complexity with a lower number of fine-tuning steps presented a challenge. The compromise led to a higher Word Error Rate (WER), indicating the impact of insufficient training steps on the model's language understanding and transcription accuracy.
- **Training Issues**: During the training process, the project encountered the errors of "IndexError - Invalid key: 3403 is out of bounds for size 0" occurred during model training", "CUDA Setup failed - Despite the availability of GPU, the setup failed" and "NotImplementedError Cannot copy out of meta tensor; no data!" hindered the fine-tuning process.
- **Fine-Tuning**: To address these challenges, certain methods such as PEFT and LORA were skipped in the fine-tuning process. Further investigation and adjustments were made to overcome errors related to CUDA setup and meta tensor copying.

Model Push to Hub

The fine-tuned model is pushed to the Hugging Face model hub for easy access and sharing.

Conclusion

- In conclusion, the Whisper-large-v3 model, despite showcasing superior accuracy, encountered challenges during training, leading to automatic crashes in Google Colab. As a strategic adjustment, the Whisper-small model was employed for training and fine-tuning with a custom dataset. It's important to note that the performance of the Whisper-small model is comparatively poorer than the Whisper-large-v3 model.
- The challenges faced during the project underscore the need for a more robust infrastructure, which we plan to address in future iterations by transitioning to cloud platforms such as AWS or Azure. This move aims to provide the necessary resources and scalability to overcome training limitations and further enhance the predictive capabilities of the Marathi Speech Recognition system.

References

- Mozilla Common Voice 11.0 Dataset
- Hugging Face Whisper-Small Model Fine-Tuning
- Hugging Face Model Hub - Marathi Speech Recognition
- Fine-Tuning Hugging Face Whisper Model
- Word Error Rate Evaluation between Large and Small Models