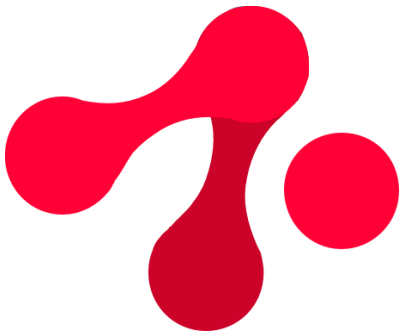

Internship Assignment for NLP(Voice AI)

Creation of Speech to Text Model

AKAIKE

30th October 2023

Goal:

- The Voice AI project aims to implement a Speech-to-Text system using the Hugging Face Whisper ASR models.
- The primary objectives include accurate transcription of Marathi audio and model fine-tuning for improved performance.

Problem Statement:

- Addressing the challenge of accurate Marathi speech transcription is crucial for applications like transcription services, voice assistants, and accessibility tools.
- Inaccurate transcription affects user experience and accessibility for Marathi speakers.

Methodology:

- The project utilizes the Hugging Face Whisper ASR models for automatic speech recognition. Fine-tuning strategies.
- PEFT (Parameter-Efficient Fine-Tuning) and LORA (Low-Rank Adaptation) technique is explored for efficient training.

Data Collection and Preprocessing:

- Common Voice Marathi dataset from Mozilla Foundation is used.
- Data preprocessing involves down-sampling audio to 16kHz, feature extraction, and tokenization using the Whisper models' feature extractor and tokenizer.

Model Architecture:

- The Whisper ASR models, specifically Whisper Small and Large versions, serve as the primary architecture and used for comparison
- PEFT and LORA adaptations are applied to improve training efficiency and adaptation to specific tasks.

Training and Fine-Tuning:

- The Seq2SeqTrainingArguments and Seq2SeqTrainer from the Hugging Face Transformers library are utilized for model training.
- Fine-tuning strategies are applied to optimize model performance.

Evaluation Metrics:

- Word Error Rate (WER) is employed as the primary metric for evaluating model performance.
- The goal is to minimize WER, ensuring accurate transcription of Marathi speech.
- Before fine tuning used provided test dataset in whisper large-v3, calculated Average WER is **73.8** and Whisper small Average WER is **93.3**

Challenges Faced:

Challenges encountered during the project include GPU memory limitations, fine-tuning difficulties, and handling large models. Strategies to overcome these challenges are discussed.

- **Storage Constraints:** The limited storage capacity in Google Colab posed a challenge, preventing the completion of additional fine-tuning steps due to insufficient space for model checkpoints and intermediate results.
- **Low GPU Resources:** The free version of Google Colab provided inadequate GPU capacity, hindering the fine-tuning of larger and more complex models. This limitation impacted the 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.

Results:

- Due to storage and GPU limitations, the Voice AI project faced challenges, leading to incomplete fine-tuning, reduced model performance, and trade-offs in model size. These constraints may result in suboptimal transcription accuracy and language understanding .
- This Fine tuning was not working as expected. But I tried my best to perform tuning.

Future Work:

Future enhancements will involve exploring additional pre-trained models, incorporating more diverse datasets, and experimenting with alternative fine-tuning techniques with adequate **GPU** and **Storage**.

Credits:

Datasets sourced from Mozilla Common Voice 11.0.

Model Tuning: Hugging Face's Whisper-Small (https://huggingface.co/openai/whisper-small)

Project Execution:

- Compare Word Error Rate of large and small models with this notebook: WER Comparison
- Fine-tune using: Fine-Tuning Process
- For inference, use: Voice AI Inference Script

Note: All code explanations has given in the google colab-notebook and connect through **Click**

THANK YOU