ITP4514 Artificial Intelligence and Machine Learning

Group Assignment

**[Voice Recognition on Cantonese]**

|  |  |  |
| --- | --- | --- |
| **Group Member** | **Student Name** | **Student Id.** |
| 1 | Leung yuk kin | 220057931 |
| 2 | LAM Yu Hin | 220060357 |
| 3 | NG Chun Kit | 220121921 |
| 4 | SHEK Hoi Fan | 220145549 |
| 5 | WONG Cho Fai | 220059807 |

Table of Contents

[Introduction / Background 3](#_Toc151320900)

[Problem Formulation 3](#_Toc151320901)

[Methodology 4](#_Toc151320902)

[Findings & Results 5](#_Toc151320903)

[Summary 7](#_Toc151320904)

[References 8](#_Toc151320905)

# Introduction / Background

Regard (Hannun, 2021), the voice recognition more common today. On figure 1, the voice recognition word error rate decreases by year. Our goal is to implement a voice recognition feature on our website that allows users to access functions through voice commands. We also want to run a voice service locally to ensure continuous availability. Moreover, we are interested in exploring how AI and deep learning can improve voice recognition performance.

一張含有 文字, 圖表, 行, 繪圖 的圖片

自動產生的描述

Figure 1 The improvement in word error rate over time

The Problem Statement would be displayed on the next section (Section 2). By then, we would discuss the methodology on Section 3. In Section 4, we would discuss the findings and results and a summary would be provided in the last section.

# Problem Formulation

Voice recognition is an important topic because it has many applications and benefits in various domains. It can enhance the accessibility and usability of devices and services for people who have difficulties with typing, reading, or seeing. Also, it can improve the efficiency and productivity of tasks that require human input or interaction.

# Methodology

The dataset from Mozilla common voice 11 ZH-HK. The dataset collects and validates voice recordings from volunteers. It is largest public voice dataset. It will use fine tune and evaluation model. Also, OpenAI Whisper model train and evaluate the dataset. It is a Transformer based encoder-decoder model. In addition, it is powerful and latest pre-train model for automatic speech recognition. The methodology will describe that include preprocess data, evaluate model, fine tune model based on (Gandhi, 2022).

1. **Preprocess Data**

The first step involves downloading the Common Voice 11 ZH-HK dataset from the Hugging Face website. Once downloaded, the audio samples in the dataset are resampled to a standardized sample rate of 16 kHz. To ensure manageable data chunks, the audio samples are split into segments no longer than 30 seconds. The next step involves converting the audio samples into Mel-Frequency Cepstrum (MFCC) representation, which is commonly used in speech and audio processing tasks. Simultaneously, the text data associated with the audio samples is transformed into vector labels. Finally, the preprocessed dataset and vector labels are saved locally, ready for further processing and modeling.

1. **Fine Tune Model**

The fine-tuning of the model utilizes the Seq2SeqTrainingArguments from the Hugging Face library. This tool enables quick setup of hyperparameters, facilitating efficient model fine-tuning. By setting the batch size to utilize the available VRAM within the GPU limit. The training process runs for 4000 steps, and model checkpoints are saved at regular intervals, such as every 1000 steps. This allows for tracking the progress and ensuring recoverability during the fine-tuning process.

1. **Evaluate Model**

The model evaluation employs the word error metric to assess the performance of the Whisper model. The Whisper model includes different versions with varying parameters, ranging from 39M to 1550M. The evaluation process involves assessing the performance of the Whisper model on different versions using subsets of the dataset. For example, evaluations are conducted using 10 and 1000 rows of the dataset. The Whisper model have different version that is contain different parameters. The parameters between 39M to 1550M. The evaluation will evaluate Whisper on different versions use 10 and 1000 row of dataset.

# Findings & Results

**Fine Tune Model**

The fine-tuning of the model was conducted using small, base, and tiny parameter sizes due to hardware limitations. The training process utilized the Common Voice 11 ZH-HK dataset. Figure 2 shows that as the model's parameters or training epochs increase, the training and validation loss improve, indicating better performance. However, Figure 3 reveals that the small parameter model had a higher word error rate (WER) compared to the other two models with the same hyperparameter settings. Interestingly, training the small model with a four times larger batch size resulted in a lower WER. These findings suggest that the choice of batch size during training significantly impacts performance, and using a higher batch size or expanding the dataset could lead to further improvements in the model's performance.

一張含有 文字, 螢幕擷取畫面, 行, 圖表 的圖片

自動產生的描述一張含有 文字, 圖表, 行, 螢幕擷取畫面 的圖片

自動產生的描述

Figure 2 Figure3

**Evaluate Model**

The evaluation of the Whisper model involved testing it on a subset of 1000 rows from the Common Voice ZH-HK dataset. Based on Figure 4, it was observed that the Hugging Face Whisper model, excluding the large-v3 version, did not perform well for the Cantonese language in the ZH-HK dataset. This suggests that the Hugging Face Whisper models without the large-v3 variant might not have been specifically trained on Cantonese. However, the PIP model from the official sources demonstrated satisfactory performance for the Cantonese language. Additionally, the fine-tuned models, trained on the Common Voice ZH-HK Cantonese dataset, outperformed the PIP version in the base and tiny parameter sizes. However, the small parameter model (with different training batch sizes) performed worse than the PIP version. These results indicate that fine-tuning the models on language-specific datasets can lead to improved performance, while the Hugging Face Whisper model may not be optimized for Cantonese training data.

一張含有 文字, 螢幕擷取畫面, 圖表, 行 的圖片

自動產生的描述Figure 4

# Summary

This project aims to implement voice recognition on a website and explore how AI and deep learning can improve its performance. The methodology involves using the Mozilla Common Voice 11 ZH-HK dataset, preprocessing the data, fine-tuning the Whisper model, and evaluating its performance. Findings suggest that training with larger parameters improves performance, but the default Whisper models did not perform well for Cantonese. Fine-tuning on language-specific datasets showed improvements, but further optimization is needed for specific languages.

***(1141 words)***

# References

Gandhi, S. (2022, 11 3). *Fine-Tune Whisper For Multilingual ASR with 🤗 Transformers*. Retrieved from huggingface: https://huggingface.co/blog/fine-tune-whisper

Hannun, A. (2021). The History of Speech Recognition to the Year 2030. *arxiv*, 10.

Source code: https://github.com/ken20020209/WhisperAIproject

**ITP4514 Group Assignment Submission Checklist**

**Before the submission of my work, I (the signed party) admit that:**

*(Please tick the corresponding box.)*

Yes No

1. I do not ***DIRECTLY COPY*** from any internet resources. ⬜ ⬜



1. I do not copy from **other students’** work. ⬜ ⬜



1. I have **summarized/paraphrased** my work from different ⬜ ⬜



kinds of materials.

1. I have fulfilled the word limitations (i.e. exceeded the ⬜ ⬜



***minimum required words***).

1. My work contains ***more than one paragraphs***. ⬜ ⬜



1. With all my works that I have referred from different ⬜ ⬜



sources, I have performed ***sufficient referencing.***

1. **If any violations on the above rules are found,** ⬜ ⬜



**I acknowledge that my marks might / would be**

**deducted in certain proportion.**



**\_\_\_\_\_\_\_19/11/2023\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**(Date) (Signature)**