運用基於生成預訓練轉換器架構的 OpenAI Whisper 多語言語音辨識引擎之台語及華語語音辨識之實作

Taiwanese/Mandarin Speech Recognition using OpenAI's Whisper Multilingual Speech Recognition Engine Based on Generative Pretrained Transformer Architecture

Yueh-Che Hsieh, Ke-ming Lyu, Ren-yuan Lyu
Department of Computer Science and Information Engineering
Chang Gung University
Taoyuan, Taiwan
m1029001@cgu.edu.tw, keming0329@gmail.com
renyuan.lyu@gmail.com

摘要

本篇論文中，我們對 OpenAI Whisper 進行台語的模型微調，使 Whisper 能夠輸出華語和台語的繁體漢字。我們使用 Hugging Face 官方所提供的 Whisper 的 Medium 和 Large-v2 模型和微調方式，並使用 CommonVoice 的台語資料集和網路上蒐集的台語連續劇影片和字幕檔共 800 小時，CER 最佳為 50.7%。我們將在後續提供我們所微調的程式碼。

Abstract

In this paper, we conducted model fine-tuning on OpenAI's Whisper for Taiwanese languages, enabling Whisper to generate both Mandarin and Taiwanese text outputs. We employed Hugging Face's official Whisper models, namely Medium and Large-v2, and their fine-tuning methodology. Additionally, we utilized the Taiwanese dataset from CommonVoice and collected around 800 hours of Taiwanese drama videos along with their subtitle files from the internet. The achieved Character Error Rate (CER) reached approximately 50.7%. We will provide the code we have fine-tuned in the subsequent updates.

關鍵字：語音辨識、台語、華語、OpenAI Whisper

Keywords: Speech recognition, Taiwanese (Minnan), Mandarin , OpenAI Whisper

1 介紹

根據台灣 2020 年人口及住宅普查，台灣人有 6,897,535 人使用台語為主要使用語言，約占總人口的 31.7%；甚至有 18,728,839 人會說台語，約占總人口的 86.0%。然而台語為非書寫語言，沒有正式的書寫方式，也鮮少有語音資料庫在網路上流通，在目前的語音辨識中難以進行建置。

台語的書寫方式目前多以漢字為主來表示，少部分未收入漢語字典則以台羅拼音表示。且部分台語音調與中文相近，亦使用相同的漢字表示。

在 2022 年 9 月 21 日，OpenAI 先前發表了 Whisper：一個使用了 680,000 小時的標記音訊，可對超過 90 種語言進行語音辨識的模型，我們實驗室在發現了 Whisper 後，立即對此模型對生活中的語音資料進行辨識正確率的統計。

在 2022 年 10 月 20 日，在 Meta 發表了使用台語連續劇語料建立的閩南語對英文的 AI 翻譯系統後，我們也開始嘗試對 Whisper 輸入我們收集的台語連續劇，並看 Whisper 對台語的辨識效果。

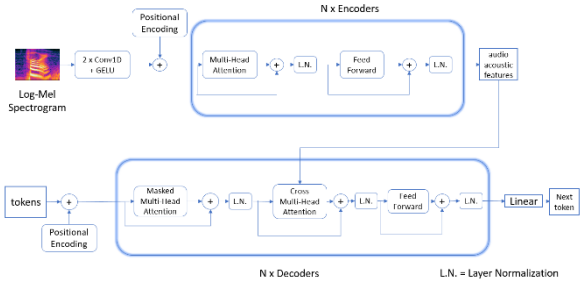
而在 2022 年 12 月 Hugging face 舉辦了 Whisper Fine-Tuning Event 並提供 Whisper 各

個模型的訓練 checkpoints 讓所有人使用不同的語言微調模型，我們在此活動中嘗試對 Whisper 使用台語進行微調，本文我們將展示我們微調的結果。

本篇論文將對 Whisper 的微調進行研究，我們使用 CommonVoice 的台語資料集和我們收集的台語連續劇對 Whisper 進行微調，嘗試使用 Whisper 對台語進行語音辨識，並輸出台語漢字或是繁體漢字。我們將使用字元錯誤率 (Character error rate, CER)作為指標。

2 模型架構

我們使用 Whisper 原論文的模型和參數進行模型微調。Whisper 使用編碼器-解碼器 Transformer(Vaswani et al., 2017)架構，圖一為 Whisper 網絡架構圖。Transformer 是一種用於自然語言處理和機器翻譯等任務的神經網絡架構。它在處理序列數據時不需要使用循環神經網絡或卷積神經網絡，而是通過自注意力機制實現了長距離依賴性的建模，Transformer 的訓練過程使用自監督學習的方法，使用遮罩語言模型預測下一個詞彙的任務進行訓練。Whisper 輸入的所有音訊都被重新取樣為 16,000 Hz，並且在 25 毫秒窗口上以 10 毫秒的步長計算出 80 通道的對數幅度 Mel 頻譜表示。



圖一：Whisper 網絡架構圖

3 實驗方法

在原本 Whisper 論文不同大小的模型對中文的辨識結果中，Medium 和 Large 的結果較優於其他三項(Tiny, Base, Small)。而台語的語法結構和中文較相近，所以在本文實驗中，我們主要對 Hugging Face 所提供的 Whisper 的 Medium 和 Large-v2 模型進行研究，對這兩個模型進行台語語音的微調，嘗試使 Whisper 能對台語進行語音辨識。

微調方式

我們使用 Whisper Fine-Tuning Event 中提供的微調方式：從 Hugging Face 下載官方提供之模型儲存點(Medium, Large-v2)，對模型儲存點使用收集的資料集進行模型微調後，對微調後模型進行辨識結果比較。

在進行台語語音的微調前，我們先對 Whisper 論文中提出的中文辨識結果進行比較，我們會先對 Hugging face 的 Medium 和 Large-v2 模型使用 Common Voice 的繁體漢字的語料庫進行微調前後的比較，一方面確認 Hugging face 提供的模型與 Whisper 論文的數據是否相近，一方面測試我們進行微調繁體漢字是否真的能提高辨識結果。

辨識模型結果的方式，我們採用 CER 作為評斷模型好壞的標準。當模型輸出的結果和標記的文本越相近時，CER 會越低，代表模型效果越好。

資料集部分

台語語音台語文字的微調，我們使用 Common Voice 的台語資料集進行模型微調與測試結果。Common Voice 資料集是用於語音技術研究和開發的大量多語言轉錄語音集合。資料集中包含 27,142 小時錄製完成的片段，其中包含 17,690 小時 108 種語言的已驗證資料。台語的部分有 120 人錄音，包含 11 小時錄製完成的片段，其中包含 3 小時的已驗證資料。在這 3 小時中，每個語音資料是以 MP3 格式儲存，標記上提供的漢字和羅馬拼音系統分別為臺灣閩南語推薦用字(部薦字)和臺灣閩南語羅馬字拼音方案(台羅)。我們將在去除台羅後，使用此資料集嘗試在輸入台語語音時產生部薦字輸出。表一提供了資料集內的部薦字、台羅和音檔頻譜圖。

臺灣閩南語推薦用字又稱部薦字，為教育部為推廣臺灣閩南語教學，及改善坊間鄉土語言教材各版本用字紊亂、紛雜不一的情形所訂定的用字。教育部亦提供「臺灣閩南語推薦用字 700 字表」列出 700 個台語建議用字和其音讀、對應華語、用例和異用字。
臺灣閩南語羅馬字拼音方案又稱台羅，為中華民國台灣教育部公布以羅馬字拼寫台語的方案。在「臺灣閩南語羅馬字拼音方案使用

手冊」中說明了台羅的音節結構是由聲母、韻母和聲調組成，並列出了所有拼音的排列方式。

台語語音繁體漢字的微調

我們從民視戲劇館 Youtube 收集了約 920 小時的台語連續劇影片，其中市井豪門 74 小時，阿不拉的三個女人 46 小時，風水世家 800 小時，並使用官方提供的字幕檔作為訓練輸入文字，該字幕檔為將台語翻譯成華語的繁體漢字。我們也將各個連續劇的 80%做為訓練資料集，剩下各 10%為驗證和測試資料集。在影片前處理我們將所有影片根據每句字幕時間點以不超過 10 秒和不超過 30 秒分割。我們將使用此資料集嘗試在輸入台語語音時直接產生相對應的繁體漢字輸出。表二舉例了我們建立的台語連續劇資料集。

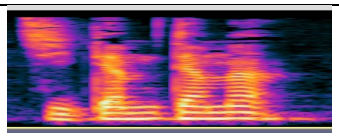
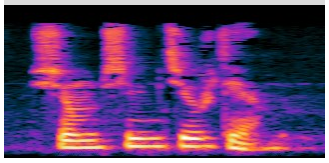
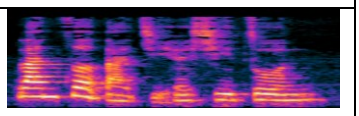
部薦字(台羅)	音檔頻譜圖
一點點仔(tsit-tiám-tiám-á)	
傷心酒店(siong-sim tsiú-tiàm)	
咱做代誌的時陣(Lán tsò tāi-tsì ê sî-tsūn)	

表 1：Common Voice nan-tw 資料集內容

檔名	文本(繁體漢字)	長度
市井_001_0094.mp3	世明 春梅 欠錢要還我還有孩子的補習費要繳	6 秒
阿不_001_0188.mp3	你會說台灣話啊 我是台灣人當然會說台灣話	4 秒
風水_001_0299.mp3	媽，先喝杯熱開水祛寒	3 秒

表 2：台語連續劇資料集內容

4 訓練結果

從表 3 中我們可以發現，在 Whisper 的原始論文中，中文的辨識結果分別為 Medium 23.2%和 Large-v2 26.8%，我們猜測這可能是對繁體漢字和簡體中文合併進行辨識的結果。而我們使用 Hugging Face 的 Whisper 和 Transformer 模組進行對繁體漢字的語音辨識，在未進行微調的 Medium 和 Large-v2 模型的辨識結果分別為 13.4%與 12.7%，在進行微調後兩者的辨識結果皆可降低至 8.9%，顯示此微調的程式在對繁體漢字的語料進行微調是有效果的。

從表 4 中我們可以發現，在使用 Common Voice 微調過後的 CER 有明顯的減少，代表著對 Whisper 的模型進行台語的微調是可以讓 Whisper 進行台語的語音辨識的。在從表 5 中觀察未微調和微調後的辨識結果時，我們也發現以下三點：

1. 在未微調的 Whisper 模型進行台語語音辨識時，可以辨識出為繁體漢字。

2. 第一句「我攏有看著」在未微調的 Whisper 模型能翻譯出是「我都看到了」。但是其他二句「大海毋驚大水」、「大鑼大鼓」的辨識結果「大害不怕大罪」、「豆河老豆河公」皆是用聲音直接轉譯的結果，無法產生有意義的句子。

3.在微調後的辨識結果，我們能看出前二句的辨識結果是完全正確的，雖然第三句「大鑼大鼓」的辨識結果「大路大股」未完全正確，但是仍比未微調的辨識正確率還要好，代表 Whisper 微調是有效果的。

從表 6 中我們可以得出，在辨識結果去除非漢字輸出如韓文、泰文和詞語重複跳針的輸出如「好好好不要拍了不要拍了慢一點慢一點慢一點…」不斷重複的「慢一點」後，使用 Large-v2 模型進行微調後對台語連續劇進行語音辨識，CER 最佳為 50.7%。我們在分析辨識結果的時候發現：

1.不論是微調 Medium 還是 Large-v2 皆容易在輸出辨識結果時，發生某些詞語重複跳針的情況，我們認為有可能是因為在收集連續劇的聲音資料時，我們並沒有進行聲音的後處理，而是直接將抓取的資料直接進行訓練和辨識

2.從表 7 中我們也可以發現，微調後的辨識結果比起未進行微調的 Whisper 更能產生正確的結果。如第一句「李有志 不義之財不可得 不倫之愛不可行」在微調後的辨識結果「李有志 不濟自財不可定 不倫不可行」就比未微調的辨識結果「余悠季 不羈季哉不叩叮不倫季艾不叩行」更為相近；「土地公 我真的不可能再有孩子了嗎」在微調後辨識結果「究竟我會不會不可能有孩子」也能辨識出大部分正確的結果。

3.雖然在微調後，模型的 CER 已經從 96.6%降低至 50.9%，要實際應用此模型仍需要進一步進行優化。CER 高的問題，我們認為的原因可能是因為：大多時候我們都直接使用繁體漢字去表示台語的對話內容，但是實際上台語轉華語也是翻譯的一種，台語對華語是一對多的，例如「骨力(kut-lát)」可翻譯成「勤勞、努力」、「擉(giáh)」可翻譯成「拿、舉起、豎起」等。這也產生了另一個問題：目前能從網路上收集的台語資料仍舊不足。就台語對華語的翻譯問題，教育部只提供了 700 字的台語華語對應，就算我們將收集的 920 小時的對應關係資料全部建立，在缺乏主要使用台語的使用者，也是一大工程，更何況不到 1000 小時的資料也難以將所有台語對華語的對應全部包含於此，這也是目前台語研究者需要面對的問題。

辨識模型	CER(%)
Whisper 論文 Medium	23.2
Whisper 論文 Large-v2	26.8
Hugging Face Medium	13.4
Hugging Face Large-v2	12.7
Medium Fine-tune	8.9
Large-v2 Fine-tune	8.9

表 3：對 Common Voice zh-tw 語料集進行 Fine-tune 之結果

辨識模型	CER(%)
Hugging Face Medium	96.6
Hugging Face Large-v2	96.7
Medium Fine-tune	50.9
Large-v2 Fine-tune	52.8

表 4：對 Common Voice nan-tw 語料集進行 Fine-tune 之結果

Common Voice 台語文字標記	未微調辨識結果	微調後辨識結果
我攏有看著	我都看到了	我攏有看著
大海毋驚大水	大害不怕大罪	大海毋驚大水
大鑼大鼓	豆河老豆河公	大路大股

表 5：台語輸出微調辨識結果

辨識模型	每句台詞長度(秒)	CER(%)
微調 Medium	10	82.6
微調 Medium	30	71.5
微調 Large-v2	10	53.8
微調 Large-v2	30	50.7

表 6： 對台語連續劇進行微調之結果

官方字幕	未微調辨識結果	微調後辨識結果
李有志 不義之財不可得 不倫之愛不可行	余悠季 不羈季哉不叩叮不倫季艾不叩行	李有志 不濟自財不可定 不倫不可行
你只會嗯 快點想辦法啊	你就 eκa 個這樣 農角園是有辦法的	你只要想 肯定有辦法
土地公 我真的不可能再有孩子了嗎	到底說我感情是不可能過敏的	究竟我會不會不可能有孩子

表 7：華語輸出微調辨識結果

5 結論

本篇論文展示了在提供 Whisper 台語語音和標記文字後進行微調後能有效的進行台語的語音辨識。我們也發現目前的評估辨識結果的 CER 並不能有效的展示出辨識效果，因為一句話能夠以多種文字排序進行翻譯，但目前台語語音和繁體漢字的對應關係資料仍然難以進行收集。我們希望未來能有精通華語和台語的研究者能收集對應關係資料，使台語有更好的評估辨識結果標準，台語使用者也能更直接的使用語音辨識等系統。

6 參考資料

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via Large-scale weak supervision. arXiv preprint arXiv:2212.04356.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., ... & Weber, G. (2019). Common voice: A massively-multilingual speech corpus. arXiv preprint arXiv:1912.06670.

Liu, C. H., Lyu, R. Y., Zhan, W. Z., Wu, J. S., Zhu, D. D., & Shi, J. L. (2019, October). 基於卷積神經網路之台語關鍵詞辨識 (Taiwanese keyword recognition using Convolutional Neural Networks). In Proceedings of the 31st Conference on Computational Linguistics and Speech Processing (ROCLING 2019) (pp. 182-191).

Chen, P. J., Tran, K., Yang, Y., Du, J., Kao, J., Chung, Y. A., ... & Lee, A. (2022). Speech-to-Speech Translation For A Real-world Unwritten Language. arXiv preprint arXiv:2211.06474.

臺灣閩南語推薦用字 700 字表 https://ws.moe.edu.tw/001/Upload/userfiles/file/iongji/700iongji_1031222.pdf

Whisper Fine-Tuning Event https://github.com/huggingface/community-events/tree/main/Whisper-fine-tuning-event

教育部網站公布「臺灣閩南語推薦用字」

https://www.dgbas.gov.tw/public/Data/762815371771.pdf

臺灣閩南語羅馬字拼音方案使用手冊

https://ws.moe.edu.tw/001/Upload/FileUpload/3677-15601/Documents/tshiutsheh.pdf