

摘要

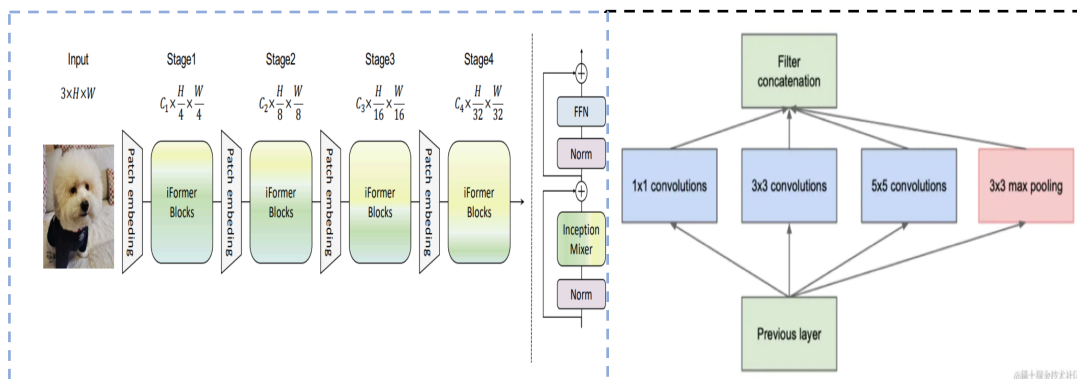
本專題研究目的是以 Transformer 模型，來實作中文語音的語音辨識。我們選擇了借鑒一個 GitHub 專案 <https://github.com/foamliu/Speech-Transformer>，提升 Transformer 模型的準確率，我們分別實作加入 convolution module、Inception module 用於提升準確率、降低模型複雜度。最後透過訓練時間、CER、參數量及複雜度來驗證改善後的模型與原先模型的優劣。

儘管現有的 ASR 技術已經取得長足的進展，通過改進已有的 Transformer 模型，我們能夠進一步提升語音辨識的準確率，同時著眼於模型的簡化和優化收斂性能。這項研究將以既有基礎為支撐，持續推動語音辨識技術的發展，並為實現更高水準的效能做出實際貢獻。

理論背景

最近的研究表示 Transformer 模型在建模遠距離依賴關係時表現出強大的能力，這得益於 self-attention 的全局計算方式，然而在捕獲更多全局信息（低頻）的同時，卻忽略了局部信息（高頻）同樣重要。

卷積因其局部計算的特點，可以有效的提取高頻信息，而 Inception 結構可以很好的處理並行結合帶來的問題，只需在進入分支前劃分合適的通道即可。同時對於不同深度的層對於高低頻資訊的需求量不同，可以控制劃分不同的通道數來適應深度，如圖一。



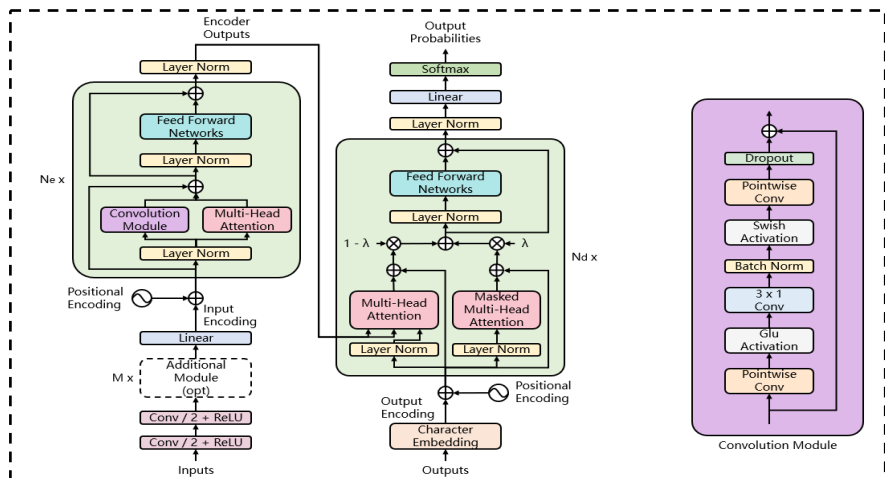
圖一、Inception Transformer 示意圖

圖二、Inception 架構

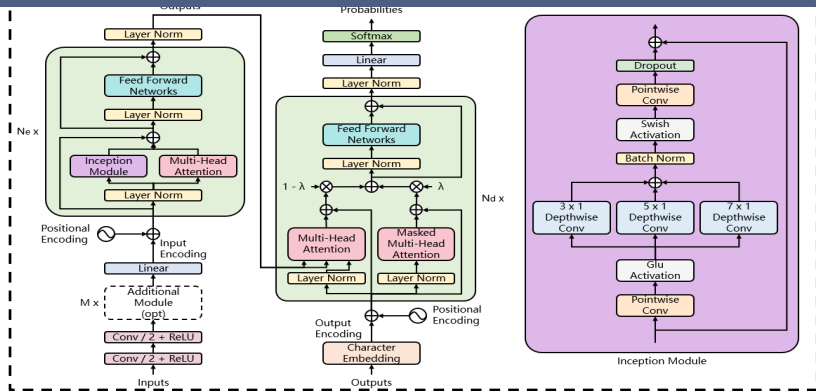
Inception 架構的核心概念是使用多尺度卷積核和池化操作結合在一起，形成一個網絡模塊，然後以這種模塊為基本單位來構建整個神經網絡結構，如圖二所示。這種模塊化的設計方式有助於更有效地構建深度神經網絡。此外形成了一個稀疏的網絡結構，這種方法既可以提高神經網絡的性能，同時也能確保計算資源的有效使用。

研究方法

下面兩張圖是我們實作 convolution module 以及 inception module 的模型架構。



圖三、convolution module for transformer



圖四、inception module for transformer

驗證方式

CER 是基於編輯距離（Edit Distance）計算的，編輯距離表示將一個字串轉換成另一個字串所需的最少編輯操作次數，如圖 15。這些編輯操作包括插入、刪除和替換字符。CER 值越低越好，表示 ASR 系統的輸出與參考文本越接近

$$CER = \frac{S+D+I}{N}$$

S 表示替換操作的次數、D 表示刪除操作的次數、I 表示插入操作的次數、N 表示參考詞數。

實驗結果與分析

	Original	Convolution Module	Inception Module
CER	0.11	0.094	0.093
Params	46.52 M	55.99 M	51.32 M
FLOPs	5.72 G	7.91 G	6.83 G

表一、實驗結果與參數表

由上表可知，加入 Convolution module 後錯誤率降低了 1.6%，而改為 Inception module 後參數量跟浮點數運算都小非常多，因為我們使用 3 個 depthwise convolution 來取代原本 3x1 的 convolution，可以捕捉不同大小的信息，不只可以使計算複雜度變小，可以在降低複雜度的同時且維持準確率。

此外我們也有實作可以丟自己的音檔做一些測試，如表二所示。可以看見其實準確率不低，有可能是受限於數據集的影響而有些部分有些微落差。其中左邊是輸出，右邊是文檔。

OUT	GT	OUT	GT
各为老师各委同学 大家好	各位老师各位同学 大家好	最后再次感谢教 授注教以及其他 专题的好伙伴门	最后再次感谢教 授助教以及其他 专题的好伙伴们
感谢教授给予的几 教能让我们在题海 中找到我们这喜欢 的题目	感谢教授给予的指 教能让我们在题海 中找到我们最喜欢 的题目	大家一提讨论深 度学习的知是互 相切从交流各自 转体的内容	大家一起讨论深 度学习的知识互 相切磋交流各自 专题的内容

表二、輸入音檔與文本的比較表

未來展望

在本文中，我們使用以 Transformer 模型為基底，並在只改動一部分模塊，而不改動任何其餘變因的情況下加強它的表現。為了得到較好的結果，我們在 Self attention 模塊並聯 convolution 模型與 Inception 模型分別測試各自的準確率、複雜度、參數量。此外未來可以在 inception 模型上優化或選用更複雜的模型使得語音辨識的模型在降低複雜度的情況下更精準。