Automatic Spoken Language Translation and Captioning for NTHU Course Videos

清華大學網路課程自動語音翻譯與字幕標記

組別：A18　　組員姓名：丁文淵、柳奕丞　　指導教授：劉奕汶

摘要

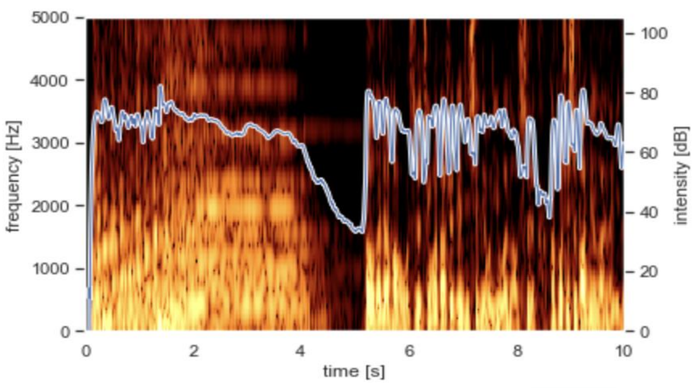
為了讓本校外國學生更方便使用「清大開放式課程」，我們設計了一套「影片翻譯系統」，其功能包含了影片下載、語音切割、翻譯以及自動生成SRT字幕標準檔。本專題之流程圖如下圖所示：

下載影片並進行轉檔 (MP4 -> WAV) → 切割音檔語句，並利用給定的語音辨識模型產生所有的中文內容 → 輸入中文內容至訓練好的模型，並翻譯成英文 → 自動生成字幕標準檔並添加字幕回原影片

我們的主要重點放在設計「語音強度偵測演算法」來辨識一句話的起點與終點時刻，還有一個Seq2Seq模型自動翻譯中文語句。
透過設計好的「語音強度偵測演算法」切割語句、紀錄語句時間點，最後將這些資訊送進語音辨識模型，並產生出該時段所講出的中文語句。接著，我們再將產生出的中文句子送進Seq2Seq模型。此模型可以做簡單的翻譯。例如：「我是清華大學OOO老師」可以轉換成「I am a teacher of college」，雖然翻譯並非100%正確，但我們可以看出此訓練後的模型已學會兩種語言的基本語法跟語意。一旦拿到英文翻譯，我們可以利用SRT字幕檔將英文字幕添加回原影片。

語音處理

Audio Intensity Detection & Segmentation



透過使用Parselmouth，我們可以呼叫函式建立音訊之時間與強度、頻率關係圖。藉此，可以判斷音訊的強度分佈以便之後進行語句切割。

Speech Recognition (Google API)

Subtitle Generation

```
2
00:00:09,150 --> 00:00:10,753
各位同學大家好
```

機器翻譯

編碼器(Encoder)：

我們將前20000個常見中文詞彙分別指定一個整數，指定完後，我們就有一句一句的index sequences了，但這些index sequences長短不一。因此，我們要統一句子的長度。首先，我們計算dataset中所有中文句子長度的平均和標準差，我們要確保所有的index sequences的長度為int(平均＋2×標準差)，若index sequence長度超過此數字，我們就要將句子截掉。若index sequence長度不到此數字，我們就要將句子進行padding。做完index sequence長度調整之後，我們就將中文的index sequences正式送進神經網路了。神經網路的第一層是Embedding層，此步驟是將每一個中文詞彙對應到一個有意義的詞向量(word embedding)。詞向量拿到之後，我們將之送進3層GRU[1]，其輸出為一個向量。此步驟的意義是將整句話的摘要濃縮在一個向量當中，我們稱此向量為context vector。此context vector將會是解碼器(Decoder)的GRU的初始狀態(initial state)。編碼器(Encoder)的架構就到此為止。

解碼器(Decoder)：

接著是處理英文句子。跟中文句子一樣，dataset裡的英文常見詞彙需要被指定一個整數。做完英文詞彙的整數指定和index sequence長度調整之後，我們就準備進入解碼器(Decoder)的Embedding層了。透過解碼器(Decoder)的Embedding層，我們就能拿到英文的詞向量(word embedding)。有了英文的詞向量(word embedding)和編碼器(Encoder)的context vector後，我們準備進入解碼器(Decoder)的GRU。我們將3層GRU的初始狀態(initial state)設定為context vector，並將我們的英文詞向量(word embedding)送進第一層GRU。當我們從第三層GRU拿到輸出之後，我們再將此輸出送進一個Dense層，Dense層的輸出會對應到一個預測的英文index，將此英文index對應到英文單字，我們就拿到我們的翻譯結果了。

專題成果

測試課程：數位聲訊分析與合成

機器翻譯 (流程圖)

Data

```
In [816]: Chinese_data[2]
Out[816]: '演示 如何 使用 ldap 執行 異步 操作'

In [817]: English_data[2]
Out[817]: 'sssss Demonstrates asynchronous operations using ldap eeeee'
```

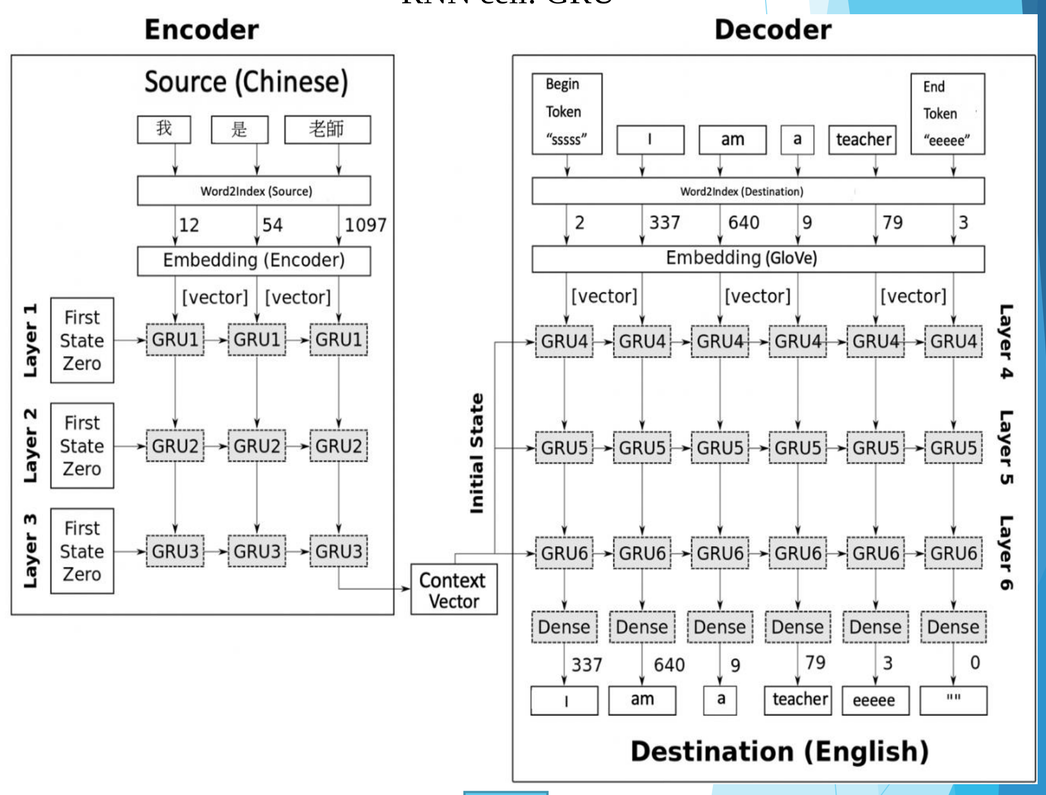
Word Embedding (300-dim for each word by GloVe)

我們使用了Stanford所研發的GloVe (Global vectors)[2]，根據許多自然語言處理實驗，GloVe 的結果通常勝過於CBOW、SVD的實驗結果。GloVe是結合自然語言處理的兩種模型(count based model、prediction based model)，透過海量的文本資料進行訓練，將相當好的詞向量整理出來。

GloVe example

```
'life': [9.3590e-02,  5.0877e-01,  1.1249e-02,.....,  3.3613e-01, -3.1116e-01, -3.9478e-01]
```

Seq2Seq model[3]
RNN cell: GRU



Translation Results

```
In [741]: translate_to_english(chinese = "這個圖是一條曲線")
          this graph is a curve
In [745]: translate_to_english(chinese = "這是常態分布")
          this is a normal distribution
```

未來可精進方向

語音處理：

1. 音訊切割演算法可改用波形的斜率來判斷。

機器翻譯：

1. 更新成Transformer模型。
2. 蒐集更多精確的中英文對照句子。
3. 基於pre-trained model，再用in-domain data 來進行fine tuning。

結論：透過這次實作的系統，我們可以有效進行語音辨識、翻譯、字幕檔產生自動化，可以有效降低人工操作時間和人力，提供英語環境供外籍學生進行學習。此外，本專題最大貢獻點在於因為網路上鮮少有特定學習科目（如微積分、普通物理、普通化學）之parallel data供訓練，所以我們透過撰寫爬蟲程式整理、歸納特定科目常用字彙的中英資料，將有助於未來其他從事相關研究、實作的研究人員。

參考文獻

[1] Z. Huang, F. Yang, F. Xu, X. Song, and K. Tsui, Convolutional Gated Recurrent Unit–Recurrent Neural Network for State-of-Charge Estimation of Lithium-Ion Batteries, IEEE Access, vol. 7, pp. 93139-93149, 2019.
[2] J. Pennington, R. Socher, and C. Manning, GloVe: Global Vectors for Word Representation, Computer Science Department, Stanford University, Stanford, CA, 2014.
[3] I. Sutskever, O. Vinyals, Q. V. Le, Sequence to Sequence Learning with Neural Networks, arXiv preprint arXiv:1409.3215, 14 Dec 2014