2019暑訓

Seq2Seq練習

（動動頭腦，試試身手）

Outline

- 資料集簡介與讀取
- 待辦事項
- 結果與討論

資料集簡介與讀取

What Kind of Data You Want to Learn

帶我走

Hi.	嗨。
Hi.	你好。
Run.	你用跑的。
Wait!	等等！
Hello!	你好。
I try.	讓我來。
I won!	我贏了。
Oh no!	不會吧。
Cheers!	乾杯!
He ran.	他跑了。
Hop in.	跳進來。
I lost.	我迷失了。
I quit.	我退出。
I'm OK.	我沒事。
Listen.	聽著。
No way!	不可能！
No way!	沒門！
Really?	你確定？
Try it.	試試吧。
We try.	我們來試試。
Why me?	為什麼是我？

當中共有19779個中英對照的句子。

待辦事項

TARGET

利用一個 seq2seq 模型,

來作英中翻譯,

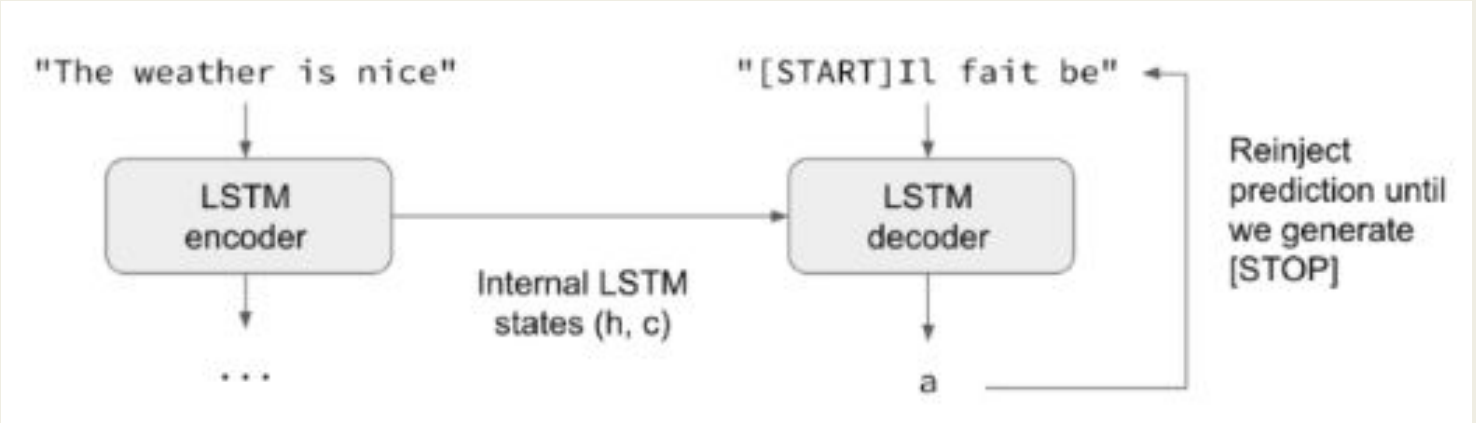
利用 LSTM ,

讓機器自我學習,

達到翻譯的功能。

Example:

『I'm fine.』-> [Seq2Seq model] -> 『我很好。』



它串連兩個LSTM隱藏層

一個隱藏層讓『input序列』擔任『編碼器』(encoder)的角色，以LSTM處理，但不管輸出，**只保留記憶狀態(State)**。

Encoder保留的**States**讓Decoder隱藏層使用，Decoder隱藏層額外再考慮**前文**，兩者綜合起來，預測下一個翻譯的字。

What You Need To Do

流程如下：

1. 讀取英/中對照檔(cmn.txt)。
2. 將raw data切割成中文句子、英文句子、中文字、英文字母。
3. 對每一個中文字、英文字母建立中文字典以及英文字典。
4. 設定 encoder_input、decoder_input對應的順序。
5. 執行編碼器(encoder)的LSTM模型，取得記憶狀態，即單字可能出現的順序，但捨棄 output，因為output是英文，我們要知道的是中文。
6. 執行解碼器(decoder)的LSTM模型，以編碼器(encoder)的記憶狀態及解碼器(decoder)的前文為input，預測解碼器目前應產出的單字。
7. 一直預測到解碼器整句結束為止。

Big Bonus

[code] 我是大伯樂ssss

```
8 #TODO: IMPORT你會用到的套件
```

```
20 #TODO:讀取資料
21 #Hint: encoding = 'utf8'
```

完成所有的藍勾勾，並將程式碼Run起來！

結果與討論

Result and Discussion