Εύρηκα

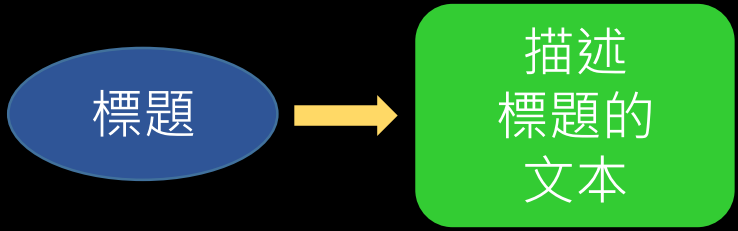
利用Transformer模型生成文本研究

專題學生：張禾牧
鄭任佑

指導老師：彭天健

研究動機

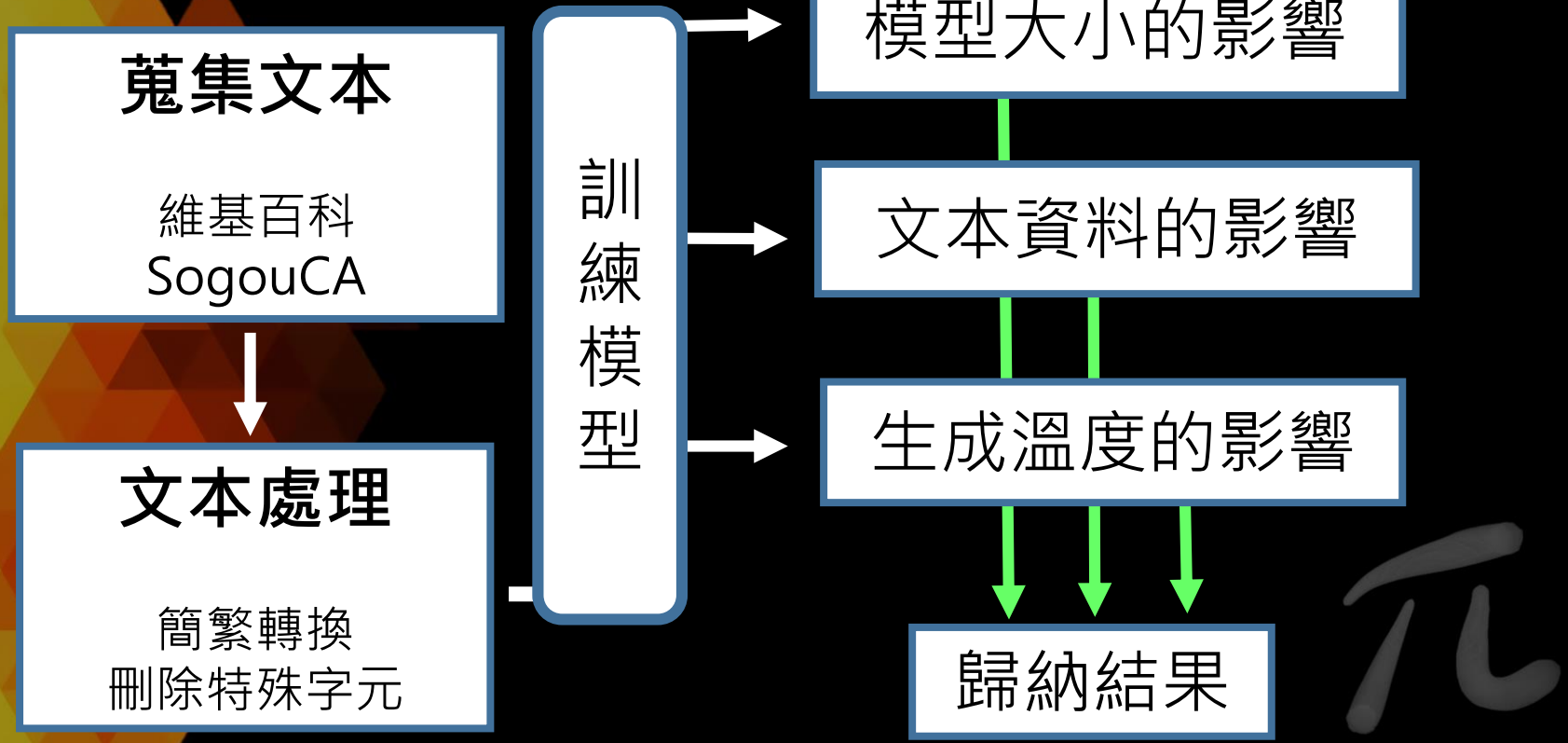
研究目的



- 生成符合指定標題的文本
- 探討不同模型超參數對成果的影響
- 探討不同訓練資料與產生內容的差異

研究過程與方法

研究架構

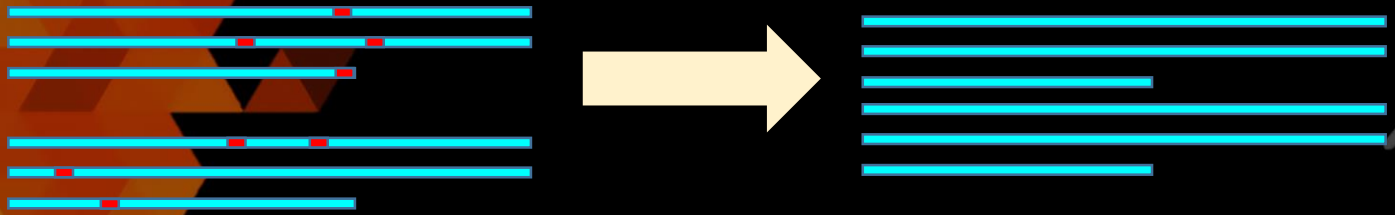

蒐集文本
維基百科
SogouCA
文本處理
簡繁轉換
刪除特殊字元
訓練模型
模型大小的影響
文本資料的影響
生成溫度的影響
歸納結果

前置處理：蒐集文本

1. 維基百科中文資料庫 2019 12 月版
2. sogou全網新聞數據 (SogouCA)

前置處理：文本處理

1. 將簡體文字轉換為繁體文字
2. 刪除無意義的字元與重複的換行



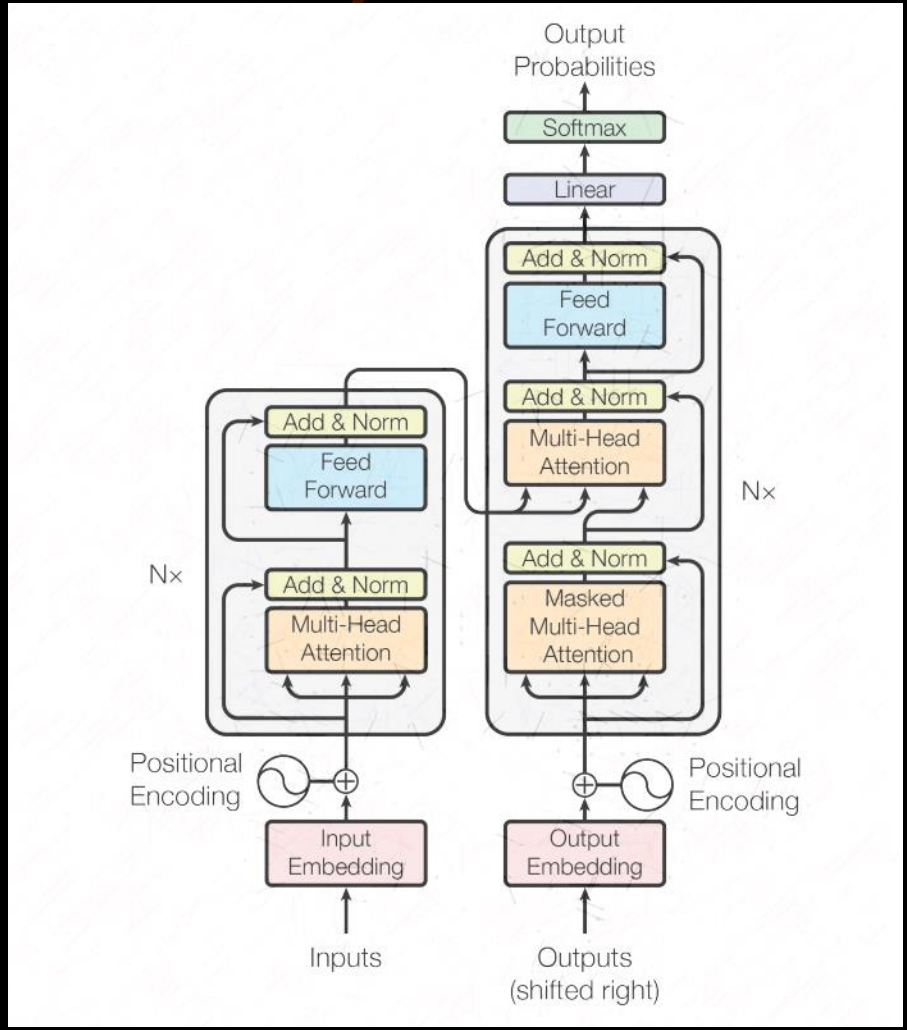
前置處理：編碼

1. 依序建立現有字之字典
2. 轉換文字list至對應的編碼

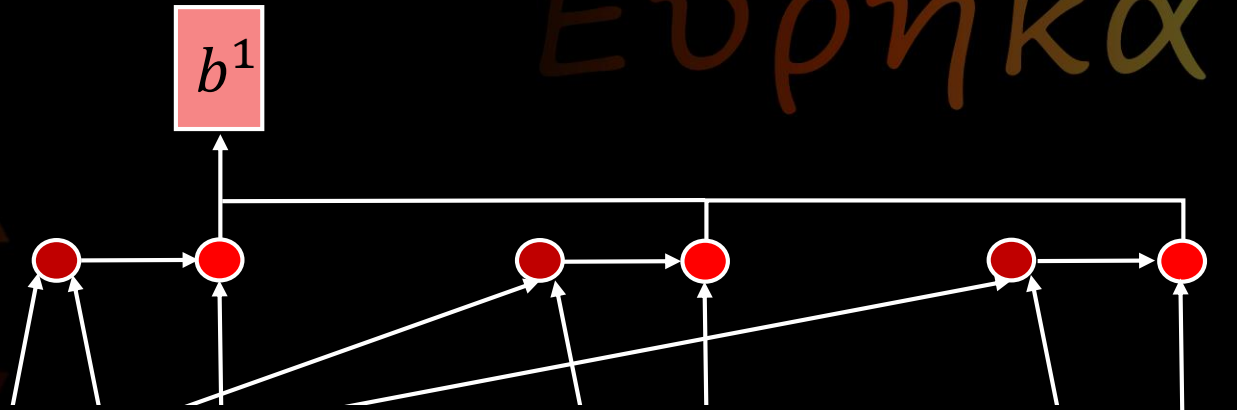
天將降大任於斯人也……

[2, 43, 55, 6, 343, 2, 43, 5, 545, …..]

Transformer



注意力機制

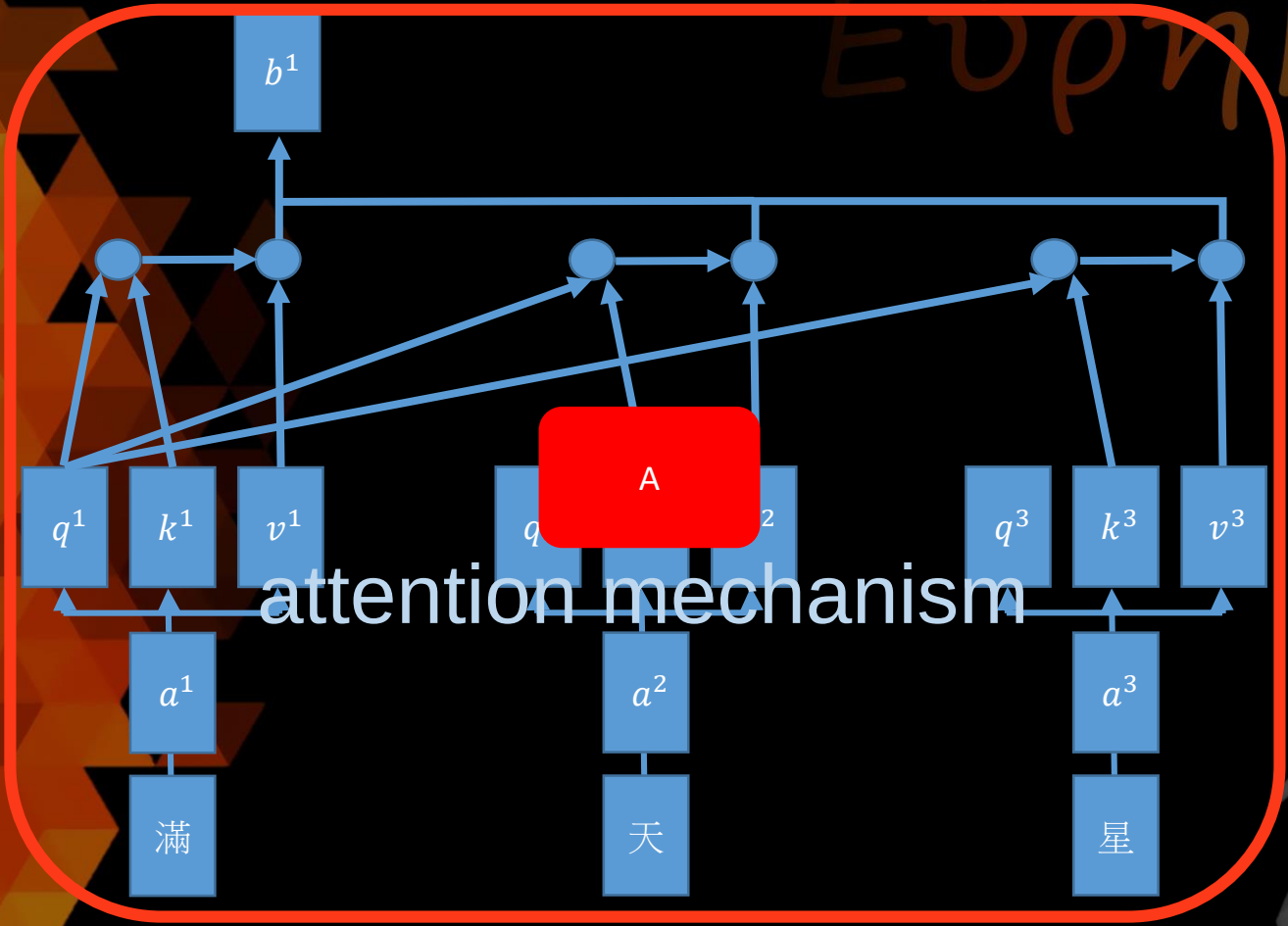


$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_K}}\right)V$$

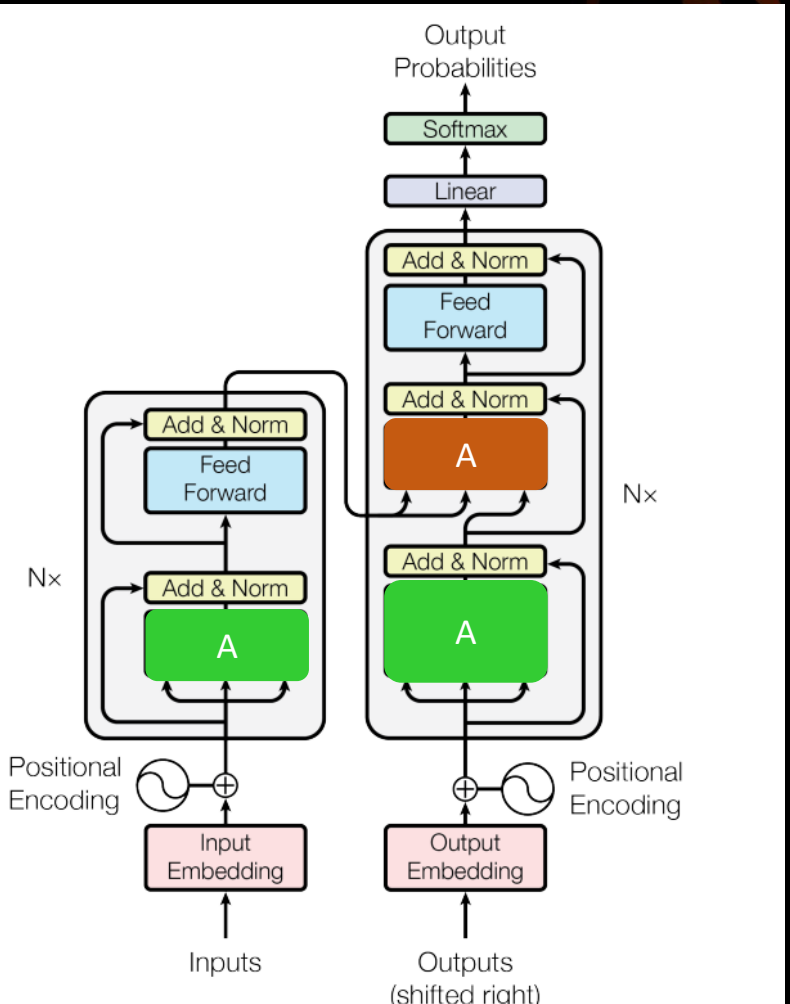
a^1 滿

a^2 天

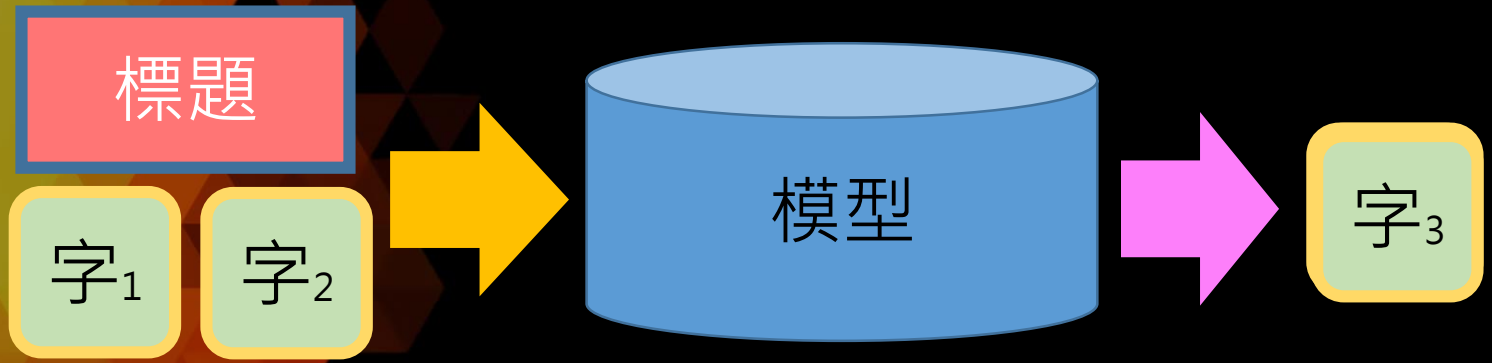
a^3 星


b^1
A
q^1
k^1
v^1
q^3
k^3
v^3
attention mechanism
a^1
a^2
a^3
滿
天
星

Transformer



文本產生過程

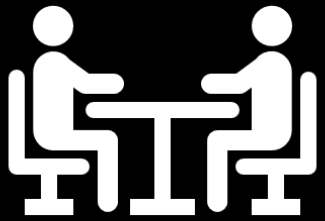


驗證：主觀評分

1.方法：十人評分

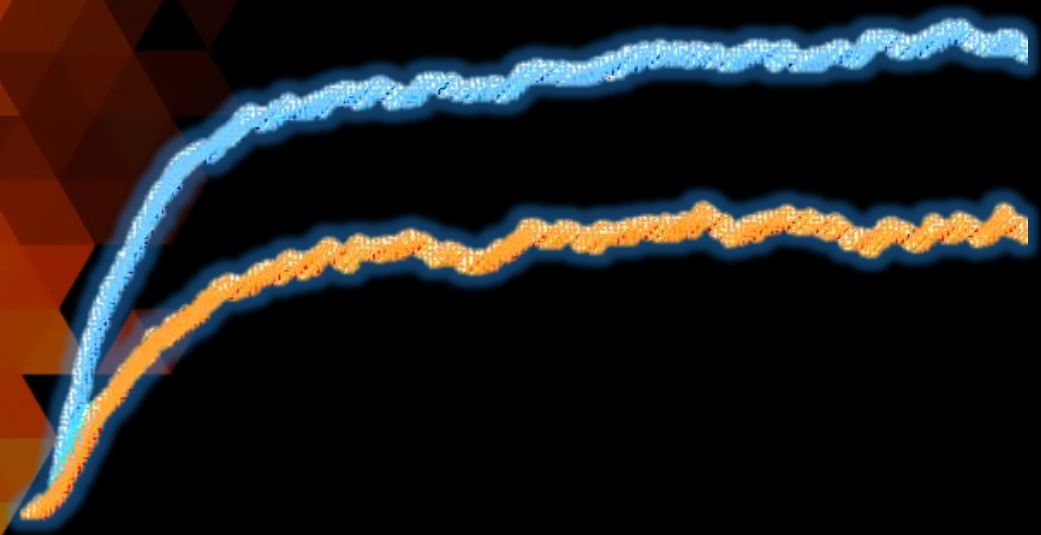
2.評分標準 (0~5分)：

語意、邏輯、用詞、標點符號與錯字、語法



驗證：客觀測試

- 模型準確度比較



研究結果與討論

文本生成示例

濱海魚類

海雅允魚，為輻鰭魚綱鱸形目刺尾魚亞目蓋刺魚科的其中一種，分佈於中西太平洋區，包括加拿大、新斯科舍海域，屬肉食性，生 活習性不明，可做為食用魚。

文本生成示例

濱海魚類

建國魚龍屬（屬名："Salinathura"）是倍鯛科下的一個品種屬，是種已滅絕捨獵魚（"S. lositoria"）、鳥腳亞目恐龍，身長151 公分，高約1汗質為弘耀型蜥巳山脈。客體長估計約1.5公尺。

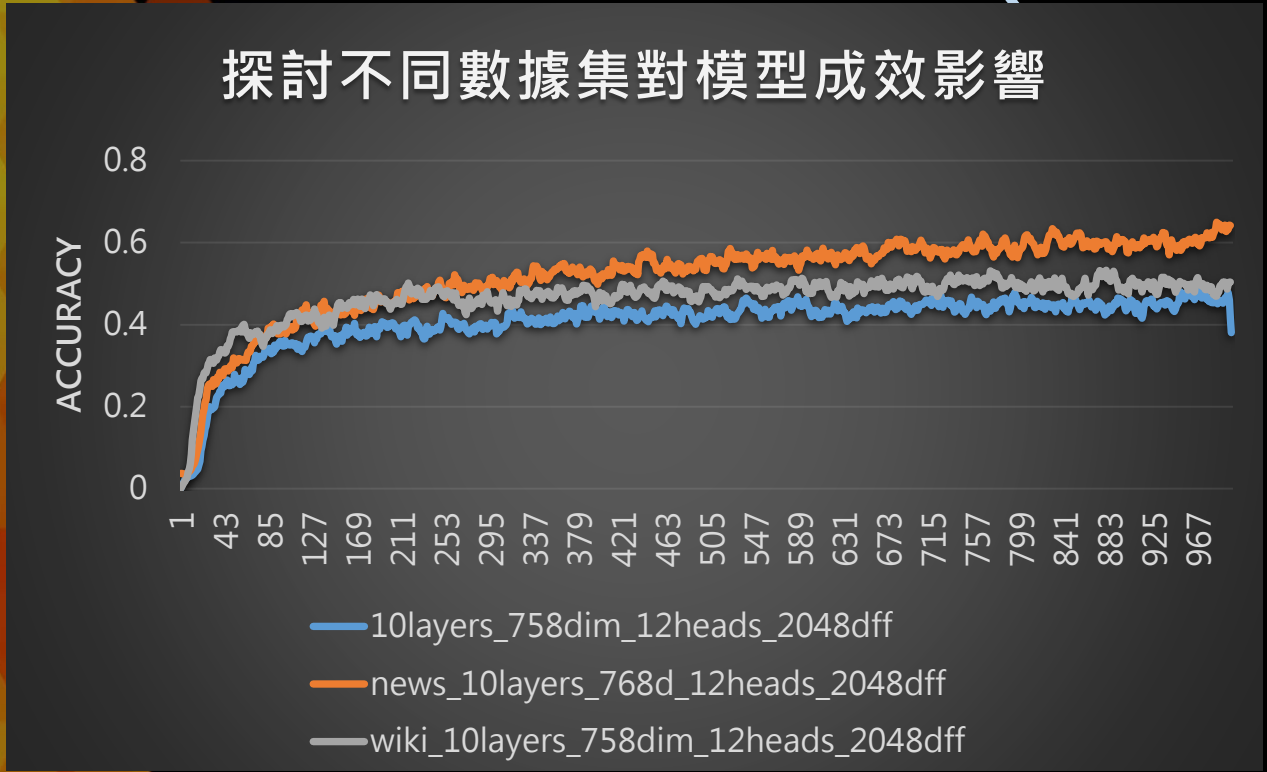
一、不同數據集比較

data	layers	dim	heads	dff
wiki	10	768	12	2048
wiki+ news_tensite	10	768	12	2048
news_tensite	10	768	12	2048

不同數據集比較：主觀評分

資料	wiki	wiki+ news_tensite	news_tensite
平均分數	2.86	2.68	2.88

不同數據集比較：客觀(準確率)測試



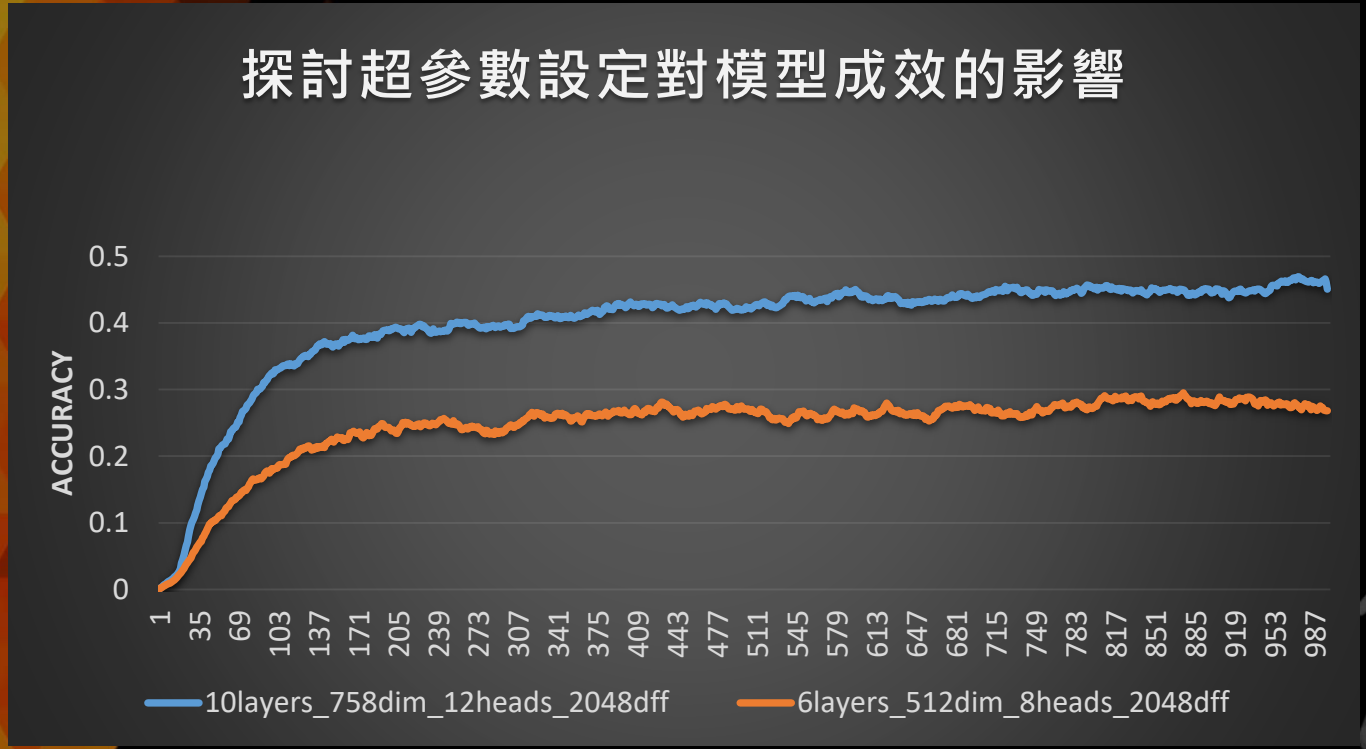
二、超參數設定比較

data	layers	dim	heads	dff
wiki+ news_tensite	10	768	12	2048
wiki+ news_tensite	6	512	8	2048

超參數設定比較：主觀評分

不同參數	未指定開頭	指定開頭
10layers_758dim_12heads_2048dff	3	3.72
6layers_512dim_8heads_2048dff	3.43	3.31

超參數設定比較：客觀(準確率)測試



三、生成溫度比較：主觀評分

模型	10layers_758dim_12heads_2048dff		
生成溫度	0.1	0.01	0.001
平均分數	0	1.48	1.23

結論

- 透過transformer模型，能生成與標題相關的文本
- 生成溫度過大或過小都會影響文本生成的結果
- 超參數的大小會影響模型的準確率
- 有差異的文本之間可能使訓練成效不因數據量增加而提升

未來展望

- 利用較大的文本及模型參數
- 加入文本摘要訓練
- 嘗試使用不同的模型與參數檢驗效果差異
- 增加較易使用的介面

參考資料

1. 淺談神經機器翻譯&用Transformer與Tensorflow2英翻中—by leemeng
   https://leemeng.tw/neural-machine-translation-with-transformer-and-tensorflow2.html

2. 進擊的 BERT：NLP 界的巨人之力與遷移學習—by leemeng
   https://leemeng.tw/attack_on_bert_transfer_learning_in_nlp.html

3 . Attention Is All You Need
   https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

研究設備及器材

1. 硬體設備 ：
   GPU：NVIDIA GeForce RTX2060

2. 軟體設備 ：
   Windows 10.0
   Python 3.7
   CUDA 10.0 & cudnn（GPU 運算技術及深層神經網路函式庫）
   tensorflow 2.0（包含tensorflow-gpu與tensorboard）
   zhconv（簡繁轉換套件）

感謝

- 專題指導老師彭天健
- 導師高君陶
- 專題好夥伴
- 父母