深度學習期末報告

108753206 李鈺祥

109753106 王中伶

中華民國 110 年 1 月

目錄

一、簡介

(一)動機

近幾年，深度學習成為非常熱門的技術，其分為文字、圖片、影音等領域，而不同領域中又擁有許多不同的應用，其中「生成」就是熱門的議題。然而，這些技術有時會被有心人士拿去使用，例如他們會利用該技術生成惡意訊息、無意義的廣告內容等，因此，若能辨認出是人寫的內容或是機器產生的內容，將省去過濾這些訊息的麻煩。

(二)目標

本專題將利用新聞內文產生標題，並合併人編寫的標題內容進而訓練分類，以判斷哪些是人或是機器所產生的內容。

二、文獻探討

(一)T5

以往能夠處理自然語言的機器學習模型，可以認為在模型上發展出易於理解、易用、泛用的知識，使模型能夠理解文本。然而，在自然語言的知識上，亦有層級之分，小至個別單字，大至整個句子甚至是通篇文章等。而這些知識常見於輔助任務的學習，例如詞向量。

近年來，遷移學習(Transfer Learning)是在自然語言處理領域獲得相當大的進步，基於擁有豐富資料訓練的預訓練模

型，使得開發人員、研究人員等使用者利用此模型進行下游任務(Downstream task)訓練，能夠取得不錯的成果。同時，這也是近年來在自然語言處理領域的發展趨勢，而遷移學習，也帶起了多樣化的方法論與實作方式。

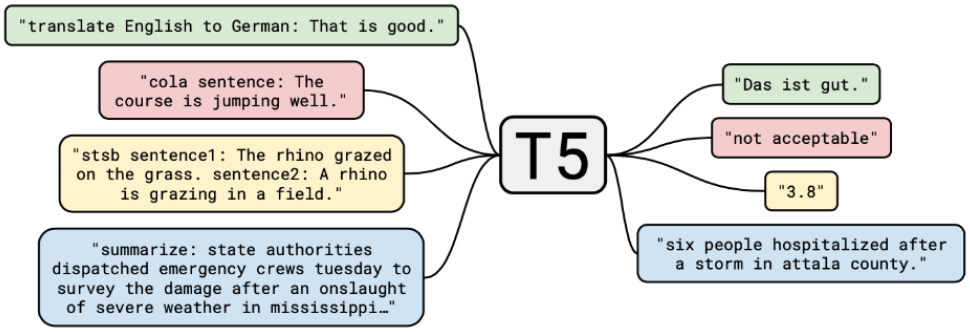


Figure 1: A diagram of our text-to-text framework. Every task we consider—including translation, question answering, and classification—is cast as feeding our model text as input and training it to generate some target text. This allows us to use the same model, loss function, hyperparameters, etc. across our diverse set of tasks. It also provides a standard testbed for the methods included in our empirical survey. "T5" refers to our model, which we dub the "**T**ext-**t**o-**T**ext **T**ransfer **T**ransformer".

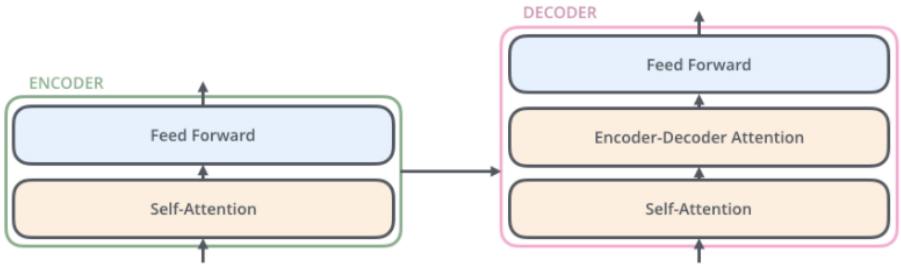
圖一、T5 模型可執行任務說明圖

T5 模型(Text-to-Text Transfer Transformer, T5)，其架構是以 Transformer 模型架構為基礎，提供了一個統一標準的使用框架，讓所有以文本為基礎的資料作為輸入，而輸出也是以文本形式為基礎，達到輸入與輸出皆是人類看得懂的自然語言，這也是 Text-to-Text 的由來。除此之外，因 T5 有一致性標準的使用框架，對於不同的自然語言處理任務，如問答系統、機器翻譯、命名實體辨識、文章摘要、情緒分析等任務，皆可以透過相同模型、目標函數、模型的訓練流程與編碼解

碼的執行過程，達到不同任務的目標。當預訓練模型並未預先訓練欲使用的學習任務時，可自行訓練新的學習任務，增加模型可處理更多樣化任務的能力。

(二)Transformer

Transformer 於 2017 年被提出，其主要概念為 Seq2Seq 模型加上 Attention 機制完成翻譯、QA、生成摘要、生成圖像描述等應用，其架構圖如下。

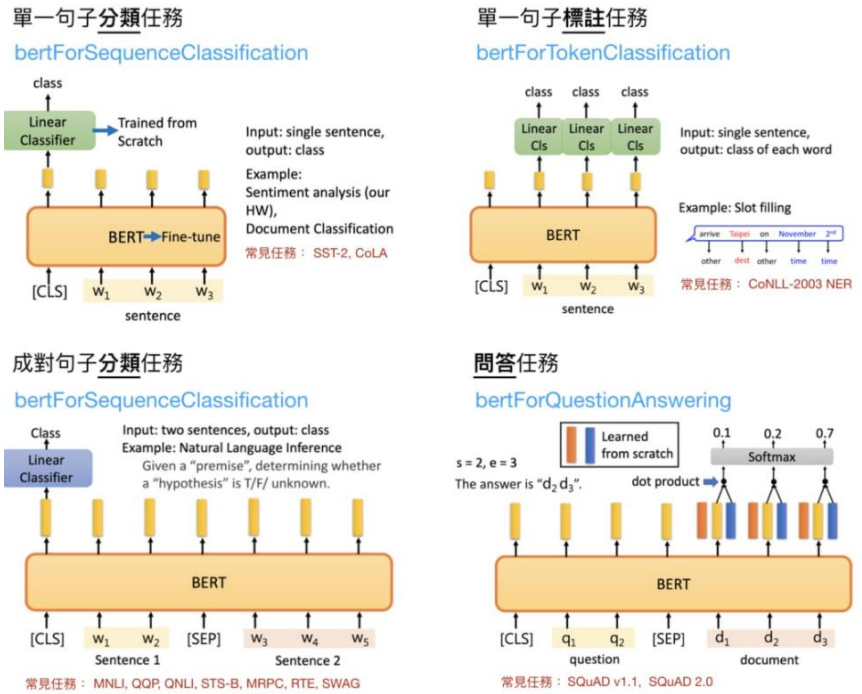


圖二、Transformer 簡易架構

提及 Seq2Seq，其模型裡的 Encoder 與 Decoder 為 RNN 架構，而 Encoder 把輸入的句子轉換成向量後，再由 Decoder 生成對應句子。然而，Seq2Seq 模型假設為能把 Encoder 輸入的內容全部壓縮成固定的語意向量，再交給 Decoder 生成，可想而知，只有一個向量是無法得到所有資訊。因此，為了解決上述問題，將 Encoder 給予 Decoder 所有輸出向量後，加上 Attention 機制，除了 Decoder 能得到更多的資訊外，其亦能專注在重要的資訊上，而這樣的架構稱為 Transformer，其架構簡易、效果良好，目前被使用在眾多文字相關應用上。

(三)BERT

BERT (Bidirectional Encoder Representations from Transformers, BERT) 於 2018 年由 Google 發表，目前也被廣泛的使用，包含分類、生成文字內容、NER 等相關任務。



圖三、fine-tuning BERT 的例子

Google 在訓練 BERT 時，進行了兩項任務，其一為克漏字填空，另外則是判斷第二句於原始文本是否與第一句相接，因此，它是一個理解上下文的語言代表模型。BERT 可被認為是 Transformer 中的 Encoder，利用大量無標註的文本所訓練的語言模型，而 BERT 擺脫以往的語言模型只能從單方向估計下個詞彙機率的問題，其利用雙向的模型獲得較多的資訊。近期，眾多人利用預先訓練好的 BERT，再對任務做 Fine-tune，其省去設計模型的成本。

三、研究方法

(一)系統架構

圖四、系統架構圖

本專案利用相同的一萬筆新聞資料，先各別訓練三種生成模型，得到各自的生成結果後，再各別進行分類的訓練，最後得到分類結果，以達成我們的目標。

(二)資料集

1. 新聞資料

資料集來源，源自於 Kaggle 資料集，News Category Dataset。此資料集內含約 20 萬篇的新聞標題、摘要、來源網址等資訊。本次實驗的新聞標題生成，需要有新聞內文方始後續的生成任務能夠順利進行，故撰寫網路爬蟲程式，對原始 20 萬篇的新聞資訊所提供的來源網址，進行新聞內容的取得。

2. 資料前處理

取得所有完整新聞內文後，記錄每篇內文的文字（word）數量。由於後續的新聞標題生成任務，在實際將資料輸入模型前，須先經過詞法分析（tokenize）與編碼轉換（encode）的處理，才能將資料輸入至模型進行訓練。

在詞法分析與編碼轉換的處理，使用 T5 預訓練模型與 BERT 預訓練模型所提供的詞法分析器（tokenizer）。由於英文經詞法分析與編碼轉換後，會使部分原本完整的英文單字，被拆解成兩個或數個子單字（sub-word），接著才會編碼成該文字在單字表中對應的編號（index）。新聞文章經轉換後，會大幅增加序列的長度。

然而，不論是 BERT 預訓練模型與 T5 預訓練模型，其最大的資料序列輸入長度為 512。因此，在實際將資料輸入至模型前，得先決定在多少的原始文章長度下，經轉換後，最大轉換後的序列長度不會超過 512。歷經反覆測試與長度的篩選，最終在原始文章長度約 155 時，在此原始長度篩選設定下的新聞文章數量約 26,068 篇，轉換後最大長度為 507。

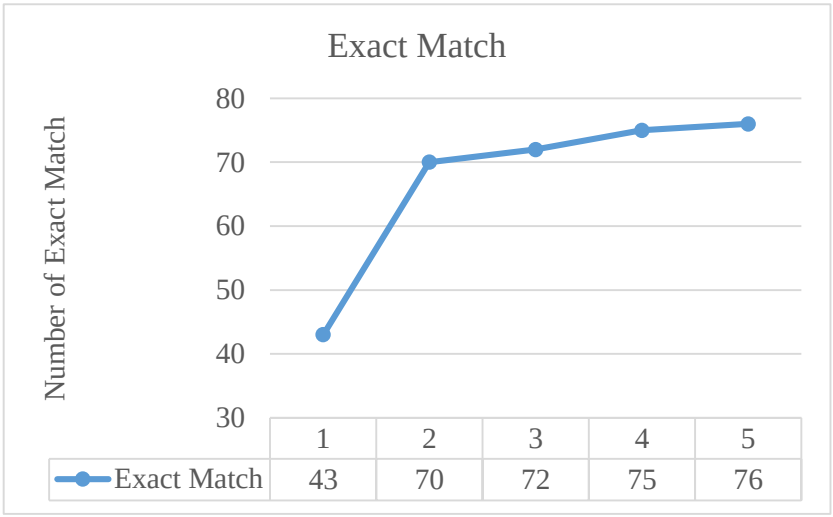
新聞標題生成的訓練集與測試集，從篩選設定下的新聞文章中，各隨機抽樣 5000 篇，其抽樣方式是採抽樣後不放回，故可避免兩者資料集有資料重複的可能。

(三)新聞標題生成

1. T5 Fine-tune

在新聞文章內容的開頭，加入任務前綴詞「headline:」的字樣，可告訴預訓練模型，在下游的學習任務要學習名為「headline:」的新任務。所有新聞內文內容經詞法分析與編碼轉換後，即可將資料輸入至模型進行訓練。此次學習任務，是做新聞標題生成，故內容部分是以新聞文章，作為學習特徵，而其新聞文章所對應的新聞標題，則是模型的預測產出。

新聞標題的生成任務，其概念與文本生成、問答系統相近或相同。在評估模型的表現上，完全符合（Exact Match）是其中一種評估方式。故評估此下游任務的學習成效，在多少個迭代學習次數下，此模型的效能會最好，並可作為之後生成的條件，其結果可參考下圖：



圖五、Exact Match 評估結果

從此圖得知，大約在迭代次數 2 之後，其完全符合的數量成長持續平緩。在後續生成提供真假新聞標題分類的任務上，模型挑選以迭代次數為 3 的模型為基礎。

2. Transformer Summarizer

以 Transformer Summarizer 模型進行新聞標題生成，其包含 5000 筆訓練資較及 5000 筆測試資料，且以 Tuple 的資料格式存取新聞內文及對應標題，以 Iterator 的方式一筆一筆讀進 Model 訓練，以下為訓練參數：

- embedding size: 500
- max_prediction_len: 20
- batch size: 100
- step_per_epoch: 300
- epochs: 5

(四)新聞標題分類

首先，先將資料標記，若是機器產生標記為 1，人所寫的標題標記為 0，接著使用 BERT Classification 進行機器產生及人編寫內容的分類，其中包含 8000 筆的訓練資料及 2000 筆的測試資料，由於 pre-trained 模型使用 bert-large-uncased，因此，先將所有資料轉為小寫，再進行訓練，並用 F1-score 與 False Negative 分析結果。

四、結果展示

(一)生成結果範例

1. 新聞內文

On race relations, President Obama is feeling optimistic. At least, that's how he comes across in an interview with NPR's Steve Inskeep, who asks if "the United States is more racially divided than it was" when he took office. "No," Obama says, "I actually think that it's probably in its day-to-day interactions less racially divided."

2. 生成結果

生成模型名稱	結果
T5 Michau	President Obama: Is America More Racially Divided Than It Was When He Takes Office?
T5 Fine-tune	Obama Says The United States Is 'Probably Less Racially Divided'
Transformer	Trump White House Leadership

(二)分類 F1-score

T5 Michau	T5 Fine-tune	Transformer
85.67%	63.99%	87.99%

本專案在進行分類時,先以多種參數找尋到最佳的分類模型,再選定一個最佳參數進行各模型的訓練,意即三個生成結果是由同一個模型進行分類。由此可推論,若生成的結果越好,其語句會與人所寫的標題越相近,因此,F1-score 表現會較差,由上表得知,T5 經過 Fine-tune 後的生成結果,比其他兩者佳。

(三)分類 False Negative

T5 Michau	T5 Fine-tune	Transformer
10.40%	35.50%	13.20%

本專案之 False Negative 意即該筆資料為機器產生，卻被判定成人所寫之標題的內容，因此，由上表得知 T5 Fine-tune 的產生結果較接近人寫的內容。

五、結論

本次期末專題使用三種不同的生成模型產生新聞標題，也許生成的結果並不與新聞有正相關，但本目標以判定是機器產生或是人所的內容進行判別。然而，由上述結果得知，T5 Fine-tune 的生成結果在本次效果最佳，意即其產生的內容與人編寫內容較為相近，因此，未來如果有進行摘要或是標題的產生，可以嘗試繼續優化該模型進行任務。

參考資料

- Colin Raffel, Noam Shazeer, etc(2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. 網址：https://arxiv.org/abs/1910.10683
- Headliner。網址：https://as-ideas.github.io/headliner/
- Huggingface。網址：https://github.com/huggingface/transformers
- Kaggle 資料集 News Category Dataset。網址：https://www.kaggle.com/rmisra/news-category-dataset
- LeeMeng。淺談神經機器翻譯&用 Transformer 與 TensorFlow 2 英翻中。網址：https://leemeng.tw/neural-machine-translation-with-transformer-and-tensorflow2.html
- LeeMeng。進擊的 BERT：NLP 界的巨人之力與遷移學習。網址：https://leemeng.tw/attack_on_bert_transfer_learning_in_nlp.html
- T5-Michau。https://huggingface.co/Michau/t5-base-en-generate-headline