師大資工系專題研究期末成果報告

計畫名稱：自然語言模型視覺化工具

執行計劃學生：朱自宇

研究期間：110 年 9 月 1 日 至 111 年 1 月底止，計 5 個月

指導教授：王科植 教授

中 華 民 國 111 年 1 月 19 日

1. 綜合資料：

（略）

1. 研究計畫內容：
2. 摘要
3. 研究動機與研究問題
4. 文獻回顧與探討
5. 研究方法與步驟
6. 成果展示與未來展望
7. 參考文獻
8. 附錄

（一）摘要

本計畫欲創建一個將自然語言模型 (BERT) 視覺化之工具。BERT的全名為 Bidirectional Encoder Representation from Transformer，來自於Transformer這個語言架構中的 Encoder部分，並使用Multi-layer與Bidirectional的設計，將每一個字以一個詞空間中的768維向量表示，並採用兩階段遷移式學習：先使用網路上的大量文本資料預訓練模型 ( Pre-training ) ，先將每個字的向量大致定型完畢，然後再依據我們所欲處理之文本任務做特定的訓練 ( Fine-tuning ) ，對於特定的詞彙採用特定的算法，讓模型特別關注我們需要處理的部分。在處理一段文本時，BERT會將文本先做斷詞處理，將一段文本切成一個個字之後，在整段文本的開頭加上一個 [CLS] token，取名自classification，顧名思義，我們可以從這個token得知這段文本的大致情況和語意；接著BERT會在句子之間與文本的結尾加上一個 [SEP] token，取名自separate，此token用來切割句子和段落。最後，BERT採用雙向訓練的方式，在訓練文本時與一般的架構不同，除了會將前面的數據陸續往後推廣相關外，也會同時將後面的數據往前相關，運用更大的關連性來處理整段句子的語意。此視覺化工具會將BERT內的資料以圖表的方式展現，以幫助自然語言分析師和資料科學家能更好的探討在此自然語言模型中，相同訓練模型對於不同文本、不同訓練模型對於相同文本、不同訓練模型對於不同文本所產生之分詞資料的差異性，並比較不同訓練模型在訓練過程中所關注的資料特性、處理資料時所作的轉換、運算等，最終能達到對於此語言模型更深一步的認識，以利於未來對於此語言模型的調整，改良，甚至簡化等工作。

（二）研究動機與研究問題

在現今的社會中，人工智慧的技術已經相當發達，科學家們已經能夠使用類神經網路訓練讓電腦做出各式各樣擬人化的行動，並且無論是在遊戲、運動還是學科等都達到十分優異的表現。但是，這些看似無所不能的電腦卻都有著一個致命性的缺陷，也是他們無法成為人類最大的阻礙—電腦無法理解人類的情感，更無法理解情感本身的存在，這也是為何現今無論人工智能機器人在各大領域表現的多麼突出，仍然時常會惹人笑話的一個主要原因。如果能使人工智能透過分析文本，定義詞彙的方式，將情感的概念以詞庫的方式讓電腦學習，或許他們就能大致的理解人類對於情感會產生的各式反應與回饋。而目前已經有人利用特定文本訓練語言模型來判斷一段文本是正向積極，還是負面的實驗，並且也達到了還不錯的正確率，但是由於語言模型的訓練機制，我們依然不能理解語言模型在訓練過程中做了甚麼處理，才能判斷出一個句子所蘊含的情感。因此，若是有一個較好的視覺化工具，能幫助科學家對於語言模型做出更詳細的分析，校正和比對，或許就能找出語言模型究竟是在哪些步驟做了哪些處理才能成功判斷一段文本的情感，並且可以使用類似的方式，將研究類推到其他領域，做出橫跨多重維度，更為複雜，使機器能更貼近人類行為舉止的實驗或是調整。為此，我們需要了解 BERT 這個語言模型的一些基本概念，才能將對於資料科學家來說相對重要或是有用的數據整合並做出視覺化，方便他們未來的處理。簡而言之，我們會需要與資料科學家或是文本分析師溝通，甚至自己先處理一些簡單的分析，試圖找出有哪些數據類型對於科學家來說有用的，或是該如何呈現才能方便日後研究的進行。

（三）文獻回顧與探討

為了幫助資料科學家們整理數據，我們首先必須瞭解BERT這個語言模型中各項數據的意義，也就是要先理解 BERT 本身的運行原理。在 Attention is all you need(<https://arxiv.org/abs/1706.03762>)中，我們可以了解到BERT 其本身最特別的自注意力機制。由於一段文本中，每個詞都必定會互相關聯，並不會互不干涉，因此 BERT 便利用了文本的此種特性，在使用文本訓練數據的過程中，在對特定的詞彙作處理時，會將重心放在某些與其較為相關的詞彙中去做處理，也因此訓練出的結果才會更加的準確。所以，在BERT訓練的過程中，對於其他詞彙的注意必定會對訓練結果有著極大程度的影響，因此我們便可以從此類機制下手，透過觀察自注意力機制前後的數據結果，我們可以得知每個詞彙對於特定詞彙可能會產生特定的影響，而這些影響與我們所關注的各類文本處理任務擁有極大的關聯性。藉由分析此機制，我們或許可以得到對於特定情感而言，某些特定的詞彙對於此類任務影響甚大，並在未來的視覺化工具中針對不同的任務，將此類特定詞特別標記出來。再者，我們也可以從自注意力機制中知道詞與詞之間的相聯性，在日後處理其他文本任務，訓練其他模型時，特別關注此類詞彙的關聯，或許能夠更進一步的提高訓練模型的準確度。

（四）研究方法與步驟

為了瞭解在BERT中，有哪些數據是資料科學家比較需要的，我們必須先了解在此模型中，定義每一個詞的768維度向量其每一維度所代表之涵義。因此，在實驗最初期，我們會嘗試去對於一些特定的自然語言任務或著已經訓練好的模型做比較，嘗試找出不同文本在相同模型下，其詞向量的特定變換。例如當一個正面的句子加上一個not變成否定句之後，是不是有特定數據會有特定的改變規律，以後觀察其他文本時，如果我們發現此類數據變化，便可推得此段文本有可能是否定句。在嘗試觀察分析特定文本之數據變換後，我們便可推廣到不同模型之間的變化，將相同的文本使用任務性質相似的不同模型去訓練，觀察他們的詞向量走向，若是走向相同，則可能代表著此類走向可能就是這種特定文本任務的大致型態，未來若是遇到相同走向的數據時，我們便可推得其應為這個領域的文本。最後，我們會去嘗試比對不同的文本在不同的模型下所產生之數據變化，將模型與文本互相整合，達到理解BERT每一項數據處理之作法與功用，以便於日後縮減，調整BERT等各類工作的進行，甚至是建立新的語言模型架構時，也可採用此類方法去計算，不僅簡略設計的時間，也可以確保數據的穩定性。

（五）成果展示與未來展望

在對於基本的正向負面文本作分析後，我們發現語意相近的句子，若是從正面句變成負面句(將一個not加至正面句之中)時，在第73維時數值會急遽變小，而在第92、164、203、235、261、285維時數值會有明顯的增加趨勢(詳情見附錄中的連結(一) )。但由於我所使用的文本( 連結(二) )並不夠大量，此項結果有可能為在此特定文本下的特定性結果。因此現在觀察出此類規律後，我將會把此項規律類推至各類，更大量的文本中去進行比對，以確認此項規律的正當性。而在相同語句但是只是加了not變為正反兩面，其向量大體的走向並沒有差距過大，這代表或許對於我所使用的這個模型( bert-base-uncased)

中，此兩段文本是非常相近的，抑或著是在某些其他分類上類型極度相似，只在影響正面或負面的維度(可能是上面所述變化的維度)有較為大的變化，當然也不排除每一維度其實都有多重含意，要將維度全部整合才能代表某一項特定類別或任務結果的這種可能。而欲驗證此類假設，我也會藉由去執行其他文本，觀察其結果來驗證這些變化是否在大部分文本下都有著相同規律，又或著只是剛好這幾組文本中數據都剛好呈現如此變化。而上述所論之觀察目前都採用畫出折線圖的方式來進行，由於一個詞為一個768維的向量，我們須將所有向量都印出來方能進行較為嚴謹不失真之比對，因此我認為折線圖為較好呈現此項概念之方式，並且由於折線圖的特性，我們可以在一張圖上面放多條線來一次比對多樣數據，也可以間接節省了比對的時間成本。未來預計會將此類工具完善，製作出能讓使用者自定義想要觀看的資料，所使用之模型，還有特定資料集的數據(例如1~100向量的平均)展現之功能設定。

（六）參考文獻

* Attention is all you need.

<https://arxiv.org/abs/1706.03762>

* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

<https://arxiv.org/abs/1810.04805>

* Transformers official website

<https://huggingface.co/transformers/>

* Ian Tenny

<https://iftenney.github.io/projects/bertology/>

* A Multiscale Visualization of Attention in the Transformer Model

<https://aclanthology.org/P19-3007.pdf>

* Embedding Projector: Interactive Visualization and Interpretation of Embeddings

<https://arxiv.org/pdf/1611.05469.pdf>

# An Effective BERT-Based Pipeline for Twitter Sentiment

<https://www.researchgate.net/publication/348095625_An_Effective_BERT-Based_Pipeline_for_Twitter_Sentiment_Analysis_A_Case_Study_in_Italian>

* 科技部 大專學生研究計畫

<https://www.most.gov.tw/folksonomy/list/2af9ad9a-1f47-450d-b5a1-2cb43de8290c?l=ch>

（七）附錄

連結(一)：簡單視覺化文本處理之結果

neg\_1~5: negative.csv中第1到第5筆資料經由 ”bert-base-uncased” 模型所處理後之 [CLS] 768維向量折線圖

pos\_1~5: positive.csv中第1到第5筆資料經由 ”bert-base-uncased” 模型所處理後之 [CLS] 768維向量折線圖

<https://drive.google.com/drive/folders/1GjJaJSz0SvNonkgXJUKR-8z6F2uKpWB4?usp=sharing>

連結(二)：訓練所使用之文本檔案

negative.csv: 負面語句之資料集

<https://drive.google.com/file/d/1Uxe1g7_OYzWL2syWvMAqztIVJUH5yAHq/view?usp=sharing>,

positive.csv: 正向語句之資料集

<https://drive.google.com/file/d/1xPaJ025rFPiRuigHa-G6JGtLu4VlIOuZ/view?usp=sharing>

連結(三)：其他訓練所使用之文本檔案

<https://drive.google.com/file/d/1N1Za4BO3bN4Iuqu5ijCxAHEgjP17Xqva/view?usp=sharing>