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 6 月底止，計 9 個月

指導教授：王科植 教授

中 華 民 國 111 年 6 月 20 日

1. 學生綜合資料：

（略）

1. 研究計畫內容：
2. 摘要
3. 研究動機與背景
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，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。BERT採用兩階段遷移式學習：先使用網路上的大量文本資料（聖經, Twitter推文等）預先訓練模型 ( Pre-training ) ，將每個字做簡單的定型、分類，然後再依據我們所欲處理之文本任務做特定的處理 ( Fine-tuning ) ，使用特定的權重和專定的算法，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，將一段文本的字元分割開來，成為一個一個單字，並在整段文本的開頭加上一個 “[CLS]” token，CLS 取名自classification，顧名思義，我們可以透過這個token的向量值和其他訊息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，SEP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，除了在處理一段句子中後面的詞彙時，會考慮到前面的單字以推斷語意之外，也會同時在處理前面的單字時預先觀察後面的詞彙，運用整段句子的關連性來整理此段文本的語意。

此視覺化工具會將BERT內的各項資料，舉凡注意力機制之權重、文本語意等以各類圖表的方式展現，試圖透過此些圖表觀察BERT中資料的屬性，以達到更加了解BERT的目標。並且，此工具預期能幫助自然語言分析師和資料科學家能更好的探討在此模型中，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屬性，以及處理資料時所作的轉換、運算等，最終能對於BERT有更深一步的認識，以利於未來對於此語言模型的調整，改良，甚至簡化。

（二）研究動機與背景

在現今的社會中，人工智慧的技術早已相當成熟，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、運動還是學科問題等，都能達到十分優異的表現與良好的正確率。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，也是這些無機體想要模擬人類時最大的阻礙—無法理解「情感」為何物，甚至，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，將情感的概念以詞彙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，儘管只是模擬，但至少在語意理解和表達上都能更接近人類。

目前已經有人利用特定文本來訓練語言模型，並且成功判斷一段文本是正向積極，還是負面悲觀的實驗。此項實驗已達到了近乎完美的正確率，但是由於BERT架構本身就是採用其他語言處理架構的編碼器來實作，因此我們依然不能很好的理解BERT本身在訓練過程中，是如何理解和歸類這段文本；對於特定字元，在文本中的前後順序等，又做了哪些處理，才能將一段文本的語意統整至預先訓練的語意宇宙中。

因此，若是有一個較好的視覺化工具，能夠呈現出BERT內部各式各樣的資料特性，或許就能幫助資料科學家做出更詳細的分析、比對，以便理解模型內部對於各項語言屬性，舉凡情緒、主附關係、動作等所應用的處理機制，進而能夠去調整，讓模型的理解能更好的貼合現今的語意空間。並且在未來，能透過這些理解以改善、完善此些機制，用以建構出更簡潔、強大的語言模型，以達到更好的語意理解能力。我們也可以將此類學習理解模式，類推至其他領域，做出能橫跨多重領域，更為複雜、全面，使機器能更貼近人類行為舉止的實驗。為此，我們必定需要了解 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或許十分重要的數據整合、比對，並以視覺化的方式方便觀察與統計。

此研究希望能以視覺化工具，整理BERT這個語言模型的各項數據以及資料屬性，以達到觀察特性、比對差距，甚至做到理解模型內部機制的任務，抑或是幫助資料科學家，語言專家等整合語言模型資料，以便於他們觀察、統計與理解BERT的內部機制。簡而言之，此研究欲透過將語言模型之資料視覺化，以便於對BERT之觀察，方便增加對於此模型之理解。

（三）相關文獻回顧與既有系統探討

為了幫助資料科學家們整理數據，在建立工具前首先必須瞭解BERT這個語言模型中各項數據的意義，也就是要先理解 BERT 本身的運行機制。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，每個詞與詞之間多少會相互關聯、影響，並不會完全分開、互不干涉，因此 BERT 便利用了人類語言的此種特性，在使用文本訓練數據的過程中，對特定的詞彙作處理時，會將重心放在某些較為重要，與其他詞彙較為相關的重點詞彙中，去做特定的審查與處理，也因此訓練出的結果才會更加的準確。

在BERT訓練的過程中，其招牌的「自注意力機制」（self-attention mechanism）必定會對訓練結果有著極大程度的影響，因此我們便可以由此下手，透過觀察引用自注意力機制前後的數據結果，我們可以得知在此機制下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，我們或許可以以更好的角度切入，了解BERT內部每一層對於注意力上的不同，便可大致推測出其實作架構中各層之間所處理工作之差異。

目前已有許多將BERT內部注意力機制視覺化之工具（詳情與範例請見參考文獻），而此些工具大都著重在呈現注意力機制的部分，顯示詞與詞之間的相對應關係。但此類工具雖然能很好的呈現注意力機制，卻無法更深一步的探討與分析其內部運行機制與原理，只能呈現出實作的結果，為此，此研究計畫會著重在呈現這些工具無法呈現之資料屬性與特性，將BERT中最原始的數據資料一一呈現，以達到從根本來觀察、從架構去分析BERT內部實際的運行方式，進而達到此研究之主題，更好的理解此語言模型。

（四）開發工具與軟/硬體需求

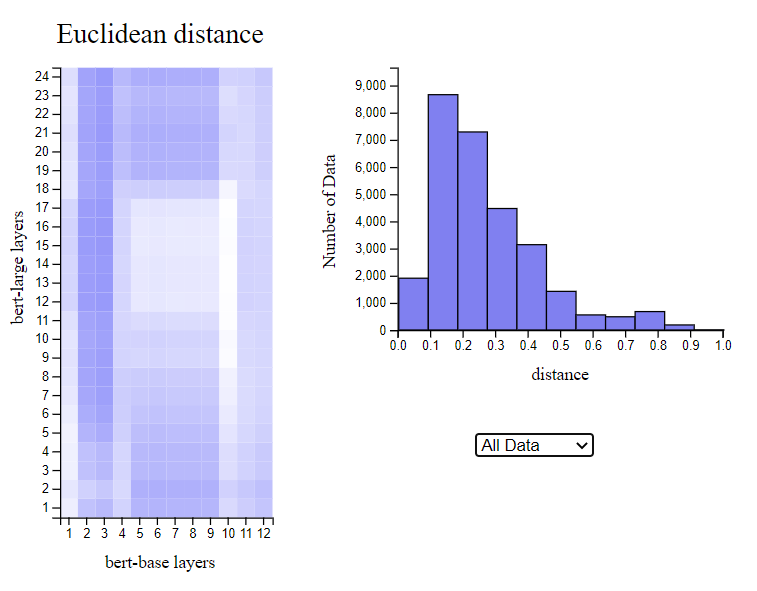
此研究使用Visual Studio Code作為主要的程式編輯器，並使用Jupyter Notebook之編輯管理功能，以Python處理資料數據，並以.csv的形式儲存，使用JavaScript的d3.js等函式庫整合資料，最終透過VS code的Live Sharing插件將視覺化之圖表以Html的方式輸出至本地網頁上。

此研究計畫對於軟體並沒有特別要求，只要擁有能夠編輯Python, JavaScript的編輯器，能夠讀取.csv檔案，並且擁有能支援網頁輸出的瀏覽器即可。在硬體方面，則建議擁有較大的記憶體空間（推薦為16 G以上），以確保一次性處理多筆資料時空間足夠，不然可能會發生記憶體不足導致設備卡頓、當機等事項。較好的CPU、GPU等也可以提升數據計算的處理速度。

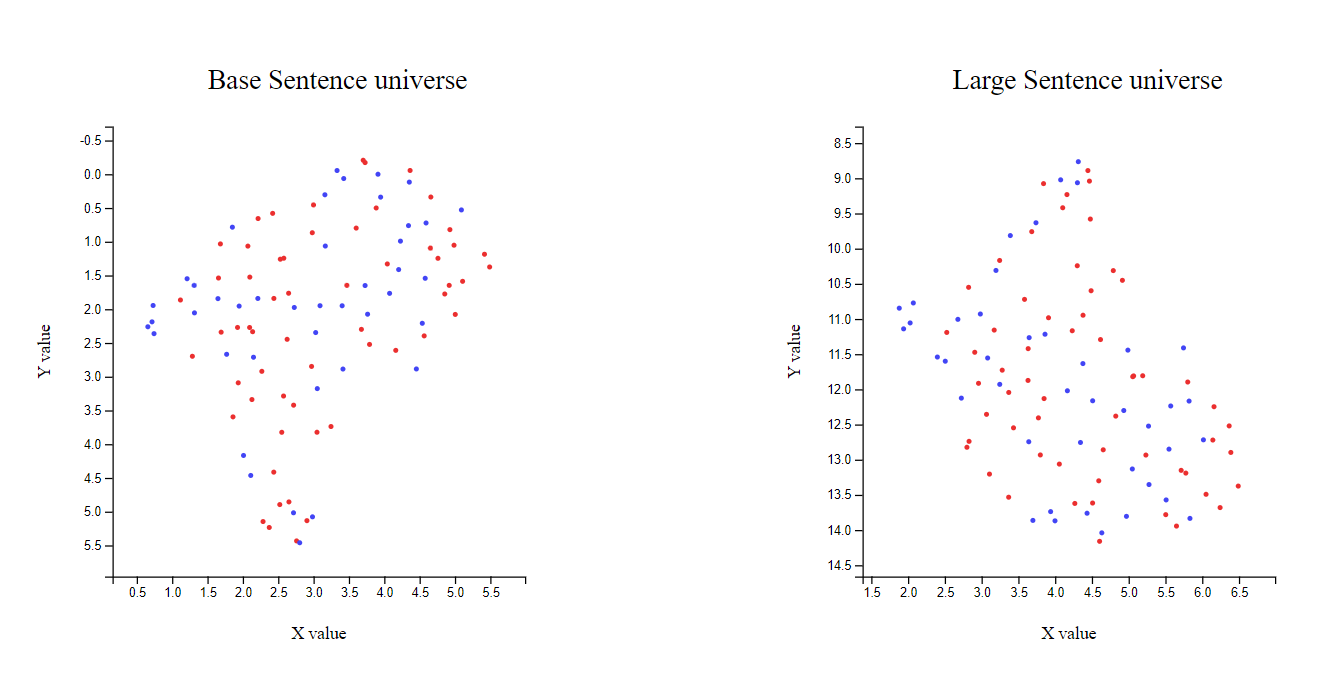
（五）系統結構、功能之展示與操作流程

欲使用此研究建構之視覺化工具，您必須先打開您的程式編輯器，並以任意形式將index.html運行在您的網頁瀏覽器上才可開始使用。

在成功運行index.html檔案之後，您可以看見幾列圖表，如下圖（一）、（二）所示：



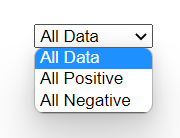
圖（一）：工具內包含之圖表示意圖



圖（二）：工具內包含之圖表示意圖

您可以透過瀏覽器的移動條將視窗移至您想要的畫面，並且透過瀏覽器的放大縮小功能將圖表依瀏覽器顯示比例放大縮小。

在此工具中，您可以移動鼠標，將其移至各圖表中的資料上（例如長條圖的長條，散佈圖的圓點等），以顯示更詳細的資訊。並且，您也可以透過頁面內提供的各類選單以切換圖表內之資料屬性顯示，如下圖（三）所示：



圖（三）：  
切換資料之  
選單式意圖

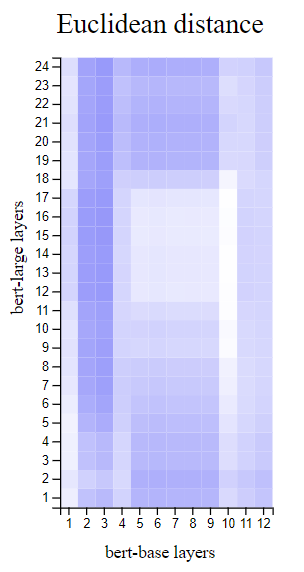
（六）系統特色

此工具著重於提供使用者各類BERT內部之資料屬性，以達到多重參考，架構比對之內容，並運用各類圖表呈現。

由此工具所提供之圖表中，我們可以看出下列特性：

1. 不同模型之間之差異

使用在工具內部之第一列圖表（如下圖（四）、（五）所示之熱度圖與長條圖），配合下拉式選單選取不同資料集，透過觀察以看出不同模型間之差異，如下列說明所示：

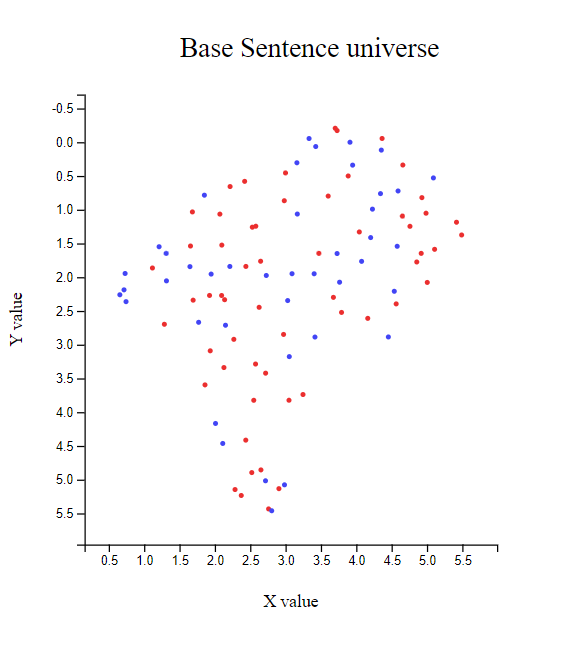


熱度圖：

在此熱度圖中，每一塊熱度圖代表著使用資料集在”bert-base-uncased”與”bert-large-uncased”下之注意力機制權重之歐式距離。顏色越深代表此資料集在bert-base的第X層與ber-large的第Y層間距離較大。此圖表中X軸的數字分別代表”bert-base-uncased”這個模型的層數（1~12層），而Y軸的數字則分別代表”bert-large-uncased”這個模型的層數（1~24層）。

藉由呈現模型間權重之距離，我們可以推斷出不同層之間的工作特性，如果兩層之間距離較遠，或許代表此兩層之處理工作較為不同。

1. 不同屬性之文本之差異



散佈圖：

在此散佈圖中，每一個點即代表著一個文本資料在BERT之語意空間中之大致位置。此圖所使用的資料集為將情緒簡略二分化為「正向」（紅色）與「負面」（藍色），將每一文本之大致語意位置呈現，以此觀察不同文本間的關係。

倘若在此圖上兩個點之距離較近，則代表在BERT之中此兩組文本是較為有關聯性的。透過觀察此散佈圖之分布，分群，我們可以簡單的看出不同屬性的文本在BERT之間的距離，由此推斷BERT的語意分類方法。

（七）各組員之貢獻

此研究計畫僅為一人所實行，所有的工作皆由40847041S朱自宇同學所包辦。

（八）後續研究/實作方向

透過此工具的觀察，我們可以在比對不同模型各層間注意力權重之圖形發現，在 bert-base-uncased模型中與bert-large-uncased之模型有少許層數之權重距離非常接近，因此我們或許可以推斷此些層數內之實作類型較為相似。後續可以使用其他類型之模型數據比較方法，透過比對多種距離計算方式，找出在大部分方式下距離最接近之模型層數，即可推斷兩者間相似程度。

再者，除了在散佈圖上提供之「正向」與「負面」之語意關係，也可以引入其他類型之資料屬性，以同樣的文本透過同樣的方式比對其各類型之屬性在BERT內語意之相近程度，以此來推斷語言模型的語言空間是以何種方式進行分類，進而了解其語言向量各維度之涵義，或是透過著重於處理特定資料屬性，優化模型之處理速度或準確率。

（九）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>

* BERT注意力機制之視覺化工具參考：

<https://github.com/jessevig/bertviz>

<https://towardsdatascience.com/deconstructing-bert-part-2-visualizing-the-inner-workings-of-attention-60a16d86b5c1>

<https://www.arxiv-vanity.com/papers/1904.02679/>

<https://www.researchgate.net/figure/Visualization-of-the-vanilla-BERT-attention-left-and-syntax-guided-self-attention_fig4_335173682>

<https://arxiv.org/pdf/2009.07053.pdf>

<https://www.topbots.com/deconstructing-bert-part-1/>