Transformer架構自然語言處理的可解釋性分析
The Explainability of Transformer in NLP

學生:尤瀚　指導教授:劉立頌 教授

摘要

現今人工智慧在自然語言處理的領域愈發成熟，隨著深度學習技術的快速發展，Transformer架構已成為自然語言處理（NLP）領域的深度學習核心模型之一。Transformer提出的注意力機制，使得其在多種NLP任務中取得了顯著的性能提升。然而，隨著模型複雜性的增加，如何理解和解釋這些黑盒模型的決策過程成為了一個重要的研究議題。

本專題旨在深入探討Transformer的注意力機制和以其為基底延伸出的BERT架構的基本原理，並分析其在自然語言處理中的應用，找出模型是如何將自然語言轉換成連續的向量，並使用視覺化工具將模型的注意力機制顯示出來，從中看出模型裡的注意力強弱。

背景

本專題使用基於Transformer的Endcoder部分的BERT架構模型。詞嵌入(Word Embedding)是BERT處理輸入的第一步，其目的是將離散的句字轉換為連續的向量表示。首先使用WordPiece 分詞方法，將句子拆分成詞或子詞，例如將「playing」拆分為「play」和「##ing」，並加上初始化標籤，用於讓模型知道句子斷點。

每個分割後的詞稱為Token ，會根據 BERT 的詞彙表映射為一個唯一的Token ID ，然後再映射到一組詞嵌入向量(Token Embedding)，加上位置嵌入(Positional Embedding)和分段嵌入（Segment Embedding），後二者與字詞在句中位置、句子所在整個輸入中的順序有關，這樣能幫助模型理解句子位置的意義。

注意力機制作為Transformer架構的核心，其作用在於理解句子內部的語意關係，一個自注意力機制（Self Attention）會將輸入的嵌入向量通過三種可學習的不同線性變換得到 Q(Query)、K(Key) 和 V (Value)三個矩陣。

自注意力機制透過Q和K來計算注意力權重並與V進行點積，透過此方法來獲取全句中詞與詞間的單個關係。多頭注意力機制(Multi-Head Attention)則是有數個「頭」，分別具有各自的K、Q、V，這樣可以從多個角度來獲取全中詞與詞間的關係。

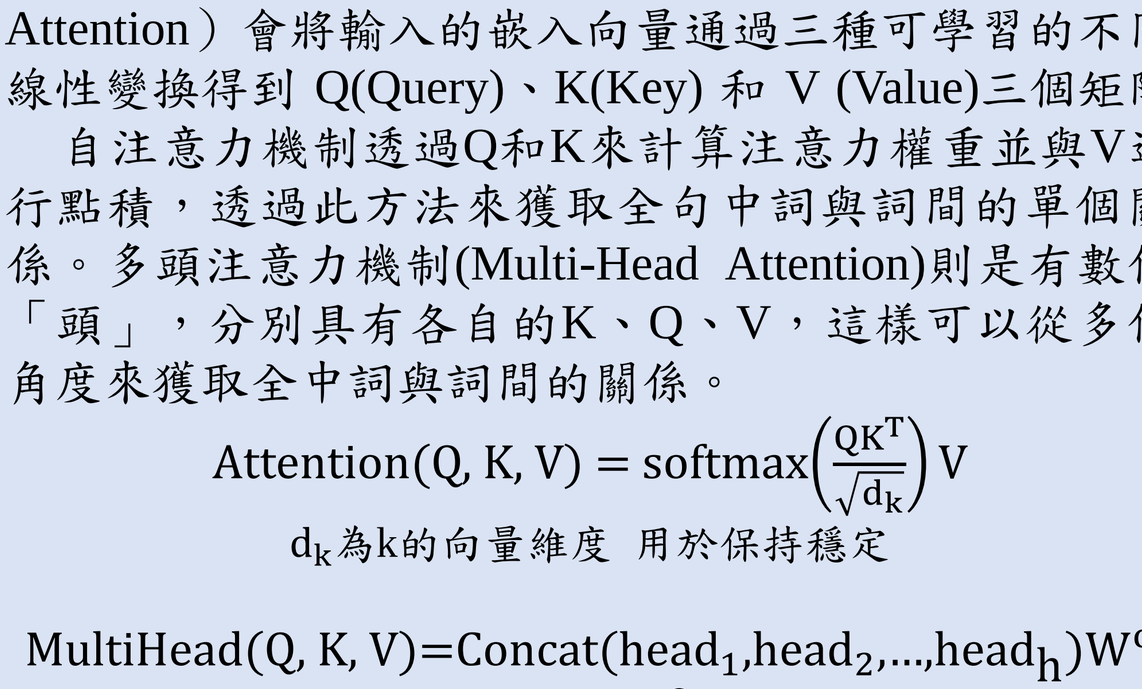
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

d_k為k的向量維度 用於保持穩定

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)W^O$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

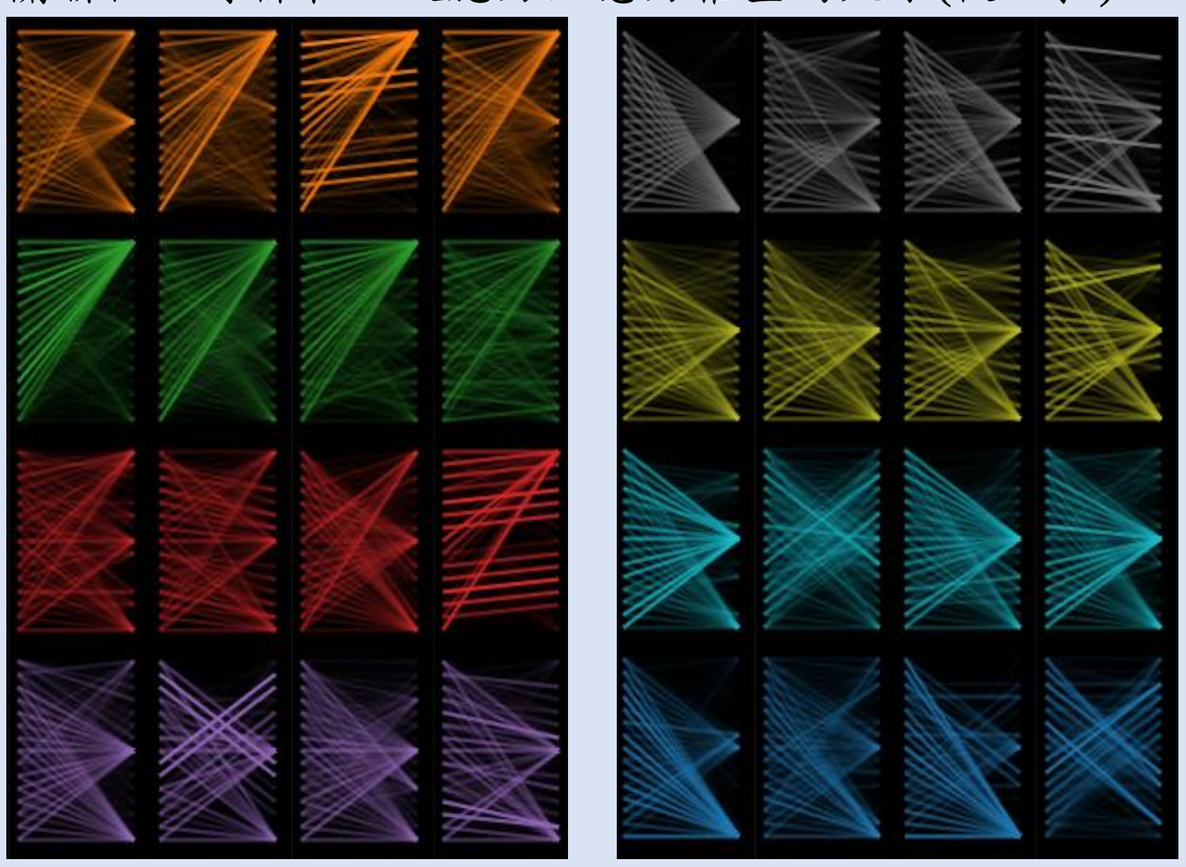
W^O為一線性變換矩陣



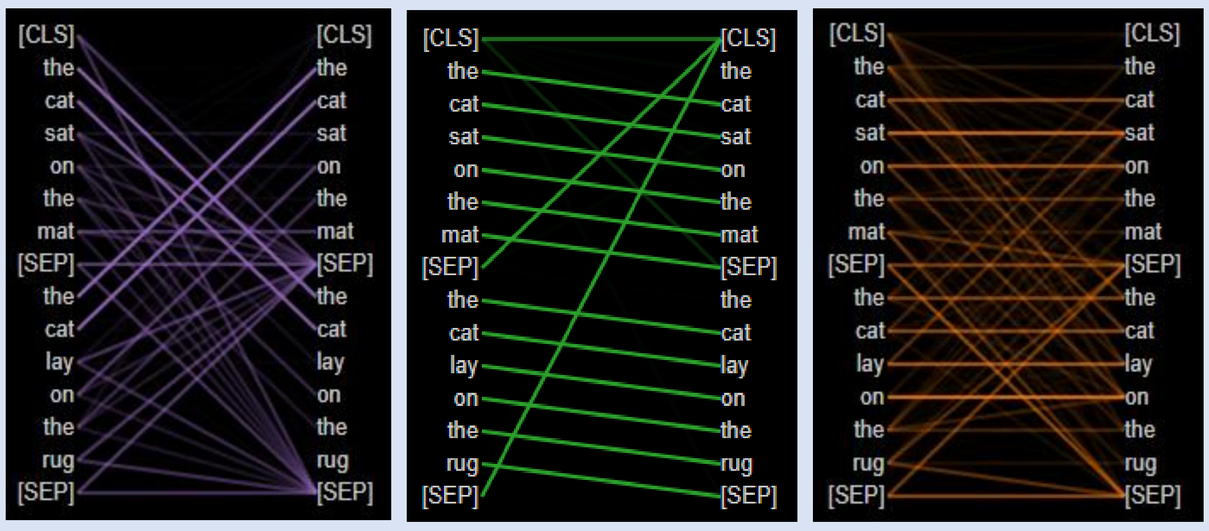
圖一、自注意力(左)和多頭注意力(右)流程圖
取自論文Attention Is All You Need

研究方法

透過輸入句子進入BERT，配合視覺化工具bertviz將模型的每一層Transformer裡的多頭注意力機制以可視化方式展示出來，更明顯的表示出每個不同的頭所注意的關聯性，線條粗細反應力注意力權重的大小(從0到1)。



圖二、將注意力可視化的部分結果



圖三、每個頭對於詞與詞之間的注意力大小

結論

這次專題探討了Transformer架構的基本原理和分析了其延伸架構BERT在自然語言處理間的詞嵌入過程與注意力機制，並了解模型如何通過計算注意力權重，使其能夠在處理語言數據時，自適應地聚焦於關鍵詞與上下文之間的關係。在可解釋性方面，強調了注意力權重的可視化對於理解模型決策過程的重要性，但是在後續訓練過程、反向傳播和結果輸出上，還有許多可解釋性的問題，期許未來在這方面能更加完善。