中　山　醫　學　大　學

醫　學　資　訊　學　系

畢業專題文件

Enhancing the Dependency Mechanism of RoBERTa

專題編號:　PRJ2020-002

專題學生:　李昱廷　郭為軒　曹仲辰

吳岳霖　林裕峰

指導教授:　張炎清 博士

中　華　民　國　一　百　一　十　年　元　月

目錄

摘要

本模型 EDM-RoBERTa (Enhancing the Dependency Mechanism of RoBERTa) 以具單頭注意力之遞迴神經網路SHA-RNN (Single-headed Attention Recurrent Neural Networks) 改良Transformer編碼器中的多頭注意力機制 (Multi-headed Attention) ，將 Boom Layer 與原始Transformer編碼器 RoBERTa 合併以重組架構，與原本分別單獨的 Transformer 與 SHA-RNN 模型相比能同時滿足長短文本序列輸入所需之短期依賴與具備 Transformer 原有的長期依賴特點。在運算過程中亦能減少計算量且提升精度及文本分類表現。

研究來源以基於 Transformer 的模型 (包含BERT、RoBERTa、XLNet、DistilBERT)、SHA-RNN 及 Transformer 中的自注意力機制為主軸進行研究，將 SHA-RNN 之 Boom Layer 改造實現注意力機制並進行高維度向量轉換，改良 Transformer 編碼器中原有的多頭注意力機制。 本研究所獲得的成果將輔助應用於情感分析、社交網路分析、聊天機器人及疾病傳播預測。

關鍵字: 情感分析、RNN、Transformer、單頭注意力、SHA-RNN、Boom Layer

一、 研究動機與研究問題

◆ 研究動機

自然語言處理是結合人工智慧和語言學領域的重要方向，注重於自然語言與電腦間的通訊互動。主要包含自然語言理解及自然語言生成，其中自然語言理解從自然語言的表達及語句中識別出該句的目的及含義。然而，機器理解過程常因段落句子組合不同而識別出不同的含義，包含一詞多義造成的理解錯誤。現行的模型訓練主要以自然語言理解 (NLU, Natural Language Understanding) 開發預訓練模型 (Pre-trained Model)，以輔助人機之間的溝通理解及後續自然語言生成 (NLG, Natural Language Generation)，因此語義特徵之提取能力尤為重要。

自然語言生成有能夠完成問答、閱讀理解、段落總結的能力。

以遞迴神經網路 (RNN, Recurrent Neural Networks) 做特徵提取，提取過程中，詞依時序讀入被分配不同的權重。然而，隨著詞與詞之距離拉遠及網路深度增加，先前被輸入詞的權重將被稀釋，造成越前面被輸入之資訊量會越來越少。以 Transformer 作為主要特徵提取模型架構，用以避免先後輸入造成的權重稀釋問題。使用多頭注意力機制針對當前預測詞將同時用到前面和後面的詞進行計算。Transformer 透過平行計算提升訓練及計算效率，RNN 則因時序複雜度太大導致計算效率低下。

Transformer 與 RNN 在單詞處理的主要差別在於時序輸入，Transformer 的無序輸入隨著段落句子及單詞組合不同而識別出不同含義，此時加入位置編碼 (Position Encoding) ，將位置編碼與詞嵌入向量內積作為輸入的嵌入向量，使詞向量包含位置訊息。輸入過程將嵌入向量透過多層的多頭注意力機制、前饋層 (Feed-Forward Neural Network)、與層標準化 (Layer Normalization)進行向量維度轉換 。輸出過程則以殘差連接連接多層的多頭注意力與前饋層。

Transformer 之特徵提取架構已相當成熟，其中注意力機制不斷經過改善以提升運算效能，以 SHA-RNN (Single-Headed Attention RNN)實作注意力達到單頭注意力的技術受到矚目。

Transformer 架構中用到多頭注意力 (Multi-Head Attention)，由於無法確認有效運算頭數，造成 Transformer 存在記憶體瓶頸與短期文本的短期依賴問題。

相較而言，SHA-RNN 在每個隱藏層中只保留單個頭的注意力，過程中有效減少多餘運算量。因此，為了提升運算效能，本研究藉 SHA-RNN 中的最佳化技術 Boom Layer 與 Transformer 中的編碼器進行修改重組 Transformer 架構執行自然語言文本分類任務。

◆ 研究問題

Transformer 的運算過程因為賦予單一句子中出現重複詞擁有相同權重，造成無法給予相對鄰近詞較大的權重，進而在自然語言理解出現問題。本研究藉由 Transformer 及 SHA-RNN 重組編碼器和解碼器架構，解決 Transformer 弱於捕獲短期文本的依賴問題，並以單頭注意力提升運算效率。

本研究使用重組後之編碼器進行更準確的自然語言文本分類任務。首先，對輸入語句以基於 Transformer 的模型(BERT、RoBERTa、XLNet、DistilBERT)進行預訓練，包含詞塊化 (Tokenization)、文本清理 (Text Cleaning) 及模型訓練 (Model Pre-training)。訓練過程以比 BERT 擁有更多訓練語料的模型 RoBERTa 透過 SHA-RNN 降低運算量，實作注意力最佳化以解決 Transformer 的短期文本依賴問題，與傳統的下映射層相比能減少整個矩陣運算量。

因此，本研究將對 Transformer 的編碼器、SHA-RNN 注意力機制與多頭注意力機制之不同、Transformer 與 SHA-RNN 的融合效果進行分析和探討。

二、文獻回顧與探討

一、Transformer

Transformer 是一種基於自注意力機制的 Seq2Seq(Sequence to Sequence)模型，用於提高神經網路機器翻譯的性能及模型訓練速度，在特定任務中的表現優於傳統的機器翻譯模型。

✧ 自注意力機制 (Self-Attention)

Transformer編碼器的輸入首先經過自注意力層，對特定單詞進行編碼並查看輸入語句的其他單詞，並輸出至前饋神經網路。

解碼器包含自注意力機制、前饋層，與用以編解碼之注意力層，協助解碼器將注意力集中於輸入語句的相關部分，避免前饋層當前解碼如圖1。

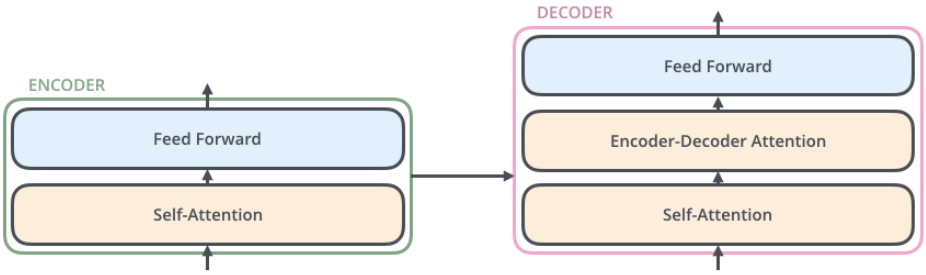


圖 1 , 編碼器與解碼器架構圖, 來源: jalammer.github.io

輸入語句的每個單詞將轉換成三個向量 (query查詢向量、key鍵向量、value值向量) ，將嵌入向量對訓練過程建立的三個矩陣 (query, key, value) 進行點乘 (Dot-Production) 產生向量。

$$A(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

透過查詢向量之點乘與各個單詞鍵向量內積得出注意力得分，為 q1·kn、q2·kn……分別除以向量維度的平方根以獲得更穩定之梯度。

透過softmax對分數進行標準化，顯示該單詞對目標單詞的相關性。將每個值向量乘以softmax分數，並加權總和就能得到自注意力輸出，如圖2。

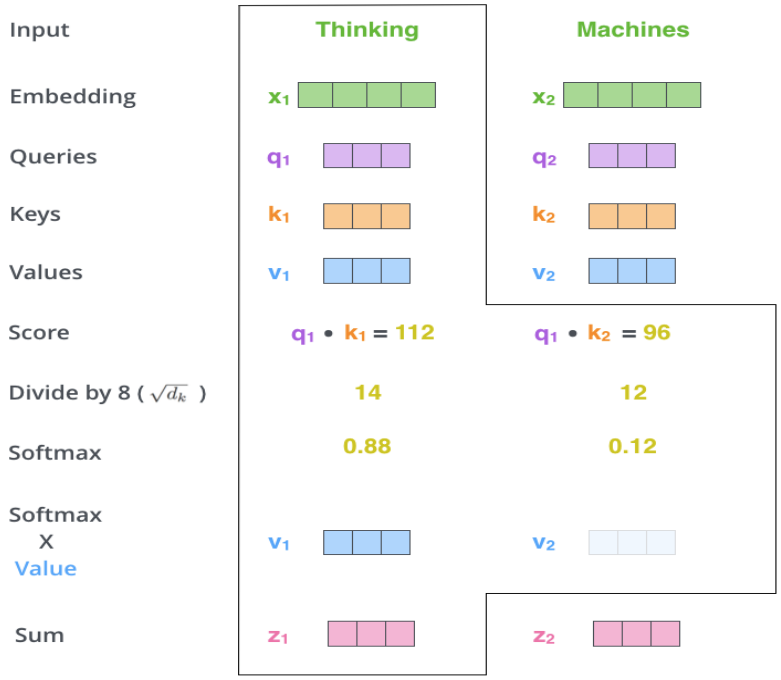


圖 2, 自注意力結構圖, 來源: jalammer.github.io

✧ 多頭注意力機制 (Multi-Head Attention)

多頭注意力為單頭注意力的原理延伸，其著重於單一查詢矩陣和多個鍵向量進行點乘並一起考慮整個輸入語句的單詞，如圖3。

頭關注的點隨著下游任務而不同，其中包含以單頭取局部的資訊或以多頭取全局的資訊。因此無論是單頭注意力機制或多頭注意力機制，在特定問題的解決方案都會被用到，如圖4。

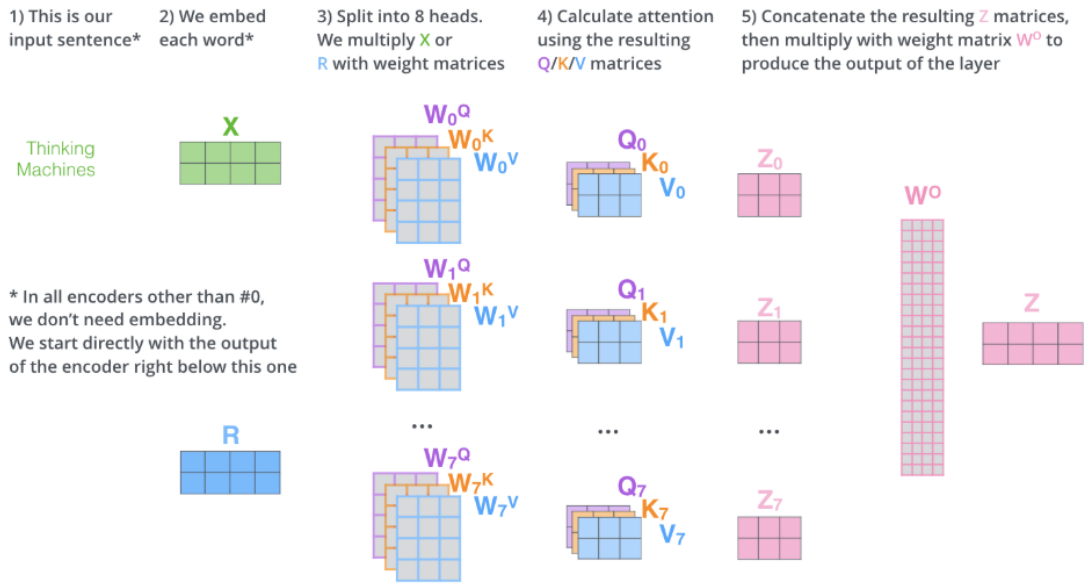


圖 3, 多頭注意力結構圖, 來源: jalammer.github.io

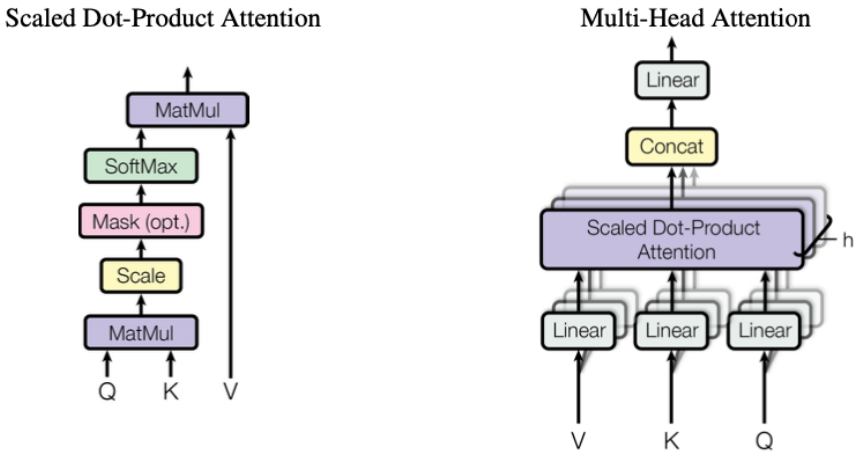


圖4, 注意力機制比較圖, 來源: Attention Is All You Need 論文

編碼過程中與傳統 RNN 相比能降低順序輸入問題，但也因無序問題衍生出位置編碼 (Position Encoding) 的必要性。在未加入位置編碼的情況下，自注意力機制會將句子中出現重複的詞賦予一樣的權重，造成相近之詞應較重要，其權重卻與另一個較遠之詞一樣重要，產生 Transformer 弱於捕獲文本中的短期依賴問題，這種對注意力的依賴會導致 Transformer 在語法敏感任務上的性能不如 RNN 模型。

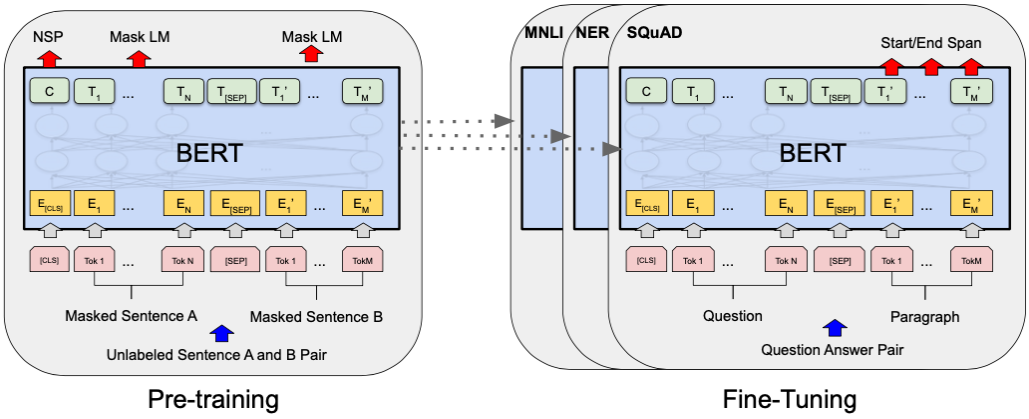
二、Bidirectional Encoder Representations from Transformer (BERT)

✧ 預訓練 (Pre-Training) 雙向 Transformer

傳統自回歸語言模型(Autoencoder)因數學定義為單向且LSTM(Long Short-Term Memory) 只能完成淺層訓練並擷取特徵，導致對於不同位置方向的單詞而言，在編碼過程看不到反向的單詞預測與理解模式，無法滿足語言處理所需的雙向共同編碼。

在訓練模式中，句子中有些單詞會依賴鄰近左右側的單詞，故僅僅從單方向編碼無法滿足上下文雙向訓練需求。

使用 Transformer進行自然語言處理任務與RNN不同之處在於能將網路做得更深，不同位置的詞能不受位置距離和方向因素而進行編碼。然而，在語言生成組合方面，縱使BERT做詞嵌入時有加入位置編碼 (Position Encoding) ，其原理是被用來與輸入嵌入求平均，因此語言組合也涉及詞序推理，並非僅需注意力機制，因此 “(Attention *isn't* all you need.) ”。



✧ Pre-training Method＃1 Masked Language Model (Masked LM)

預訓練即將大量未標註之數據進行無監督學習，使其學習語法結構、解讀語義，以完成後續有效之自然語言任務。

Masked LM 於預訓練前對訓練集中的文本序列隨機遮蔽 15% 的單詞，而非如以往將每個詞都預測一次。最後，損失函數只計算被遮蔽之詞塊 (token)，其中被遮蔽的15%中有10%被替換成其他單詞，另外 10% 不替換，剩餘 80% 被替換為 [MASK] 。

預訓練過程中，模型將猜測 15% 中的所有詞塊，對每個詞均計算損失與注意力機制，易造成當 [MASK] 出現過多，如此將影響模型收斂速度，甚至導致比傳統RNN (LSTM)模型從左至右的模型更慢。

✧ Pre-training Method ＃２: Next Sentence Prediction (NSP)

判斷第二個語句在原始文本中是否為第一句子之後續語句， RoBERTa 模型證明其視為無效之訓練方法。

✧ Transfer Learning on NLP

以語言模型預訓練方法訓練出對自然語言有相當程度理解之語言模型，將其用以做特徵擷取並針對下游任務進行微調模型。BERT能同時完成無監督學習和監督式微調。

三、Sequence to Sequence (Seq2Seq)

✧ 編碼器與解碼器

Seq2Seq 模型主要由編碼器與解碼器兩個 RNN 組成，編碼器負責將輸入序列編碼轉換成中間向量 (Context Vector)，解碼器再根據中間向量轉換成文字輸出。在預測的過程中，當前字詞的預測不僅取決於前面已翻譯的字詞，亦考慮原始輸入。

運作過程中，編碼器最後時間神經元的隱藏層輸出到解碼器的第一個神經元，透過激勵函數與 Softmax 層篩選出機率最大者做為下一個神經元的輸入，如圖5。

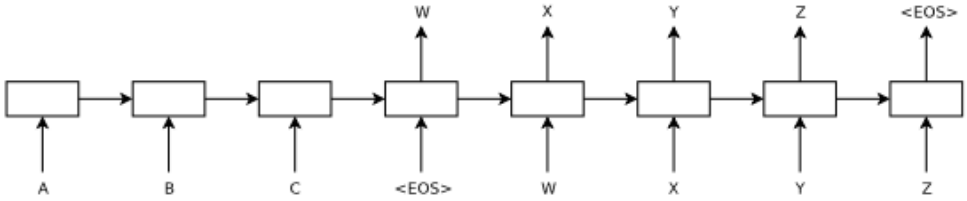
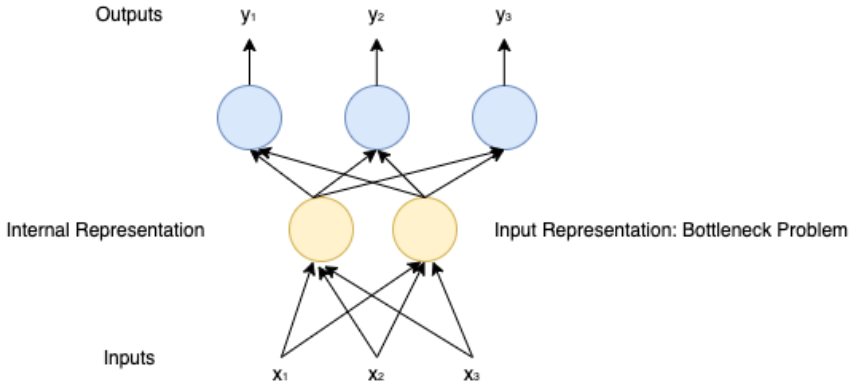


圖5, Seq2Seq模型結構圖, 來源: Seq2Seq論文

問題出現於中間向量，在編碼器以最後一個神經元進行轉換時，依序由左到右讀取資訊，但中間向量仍為固定維度的向量。導致轉換後的向量無法涵蓋所有輸入序列的訊息，先被輸入之重要訊息將在轉換後權重降低甚至消失。

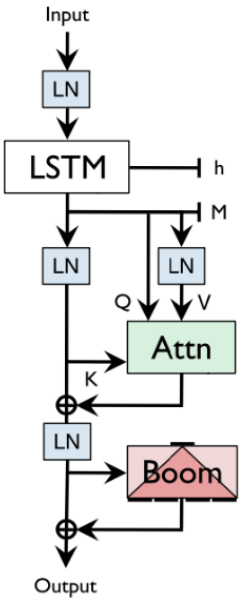


Context Vector之Bottleneck問題

四、Single-headed Attention RNN (SHA-RNN)

✧ 單頭注意力 (Single Headed Attention)

Transformer 模型建立於無序基礎，並只透過注意力機制完成訓練，但每層網路都有數十個注意力頭 (Attention Heads)，運算過程中因無法得知有效頭數而耗費多餘的運算資源。相較而言，SHA-RNN的注意力機制只保留一個頭完成向量之注意力乘積與得分。



✧ 結構變更

主要以 Transformer 之自注意力機制為基礎修改，四層結構中，每層先以 LSTM網路層生成，進行層標準化 (Layer Normalization) 後連接注意力機制，實際上為8層網路。

結構變更後與 Transformer不同之處在於本模型只對Q做全連接層，並以sigmoid產生Q, K。然而，過程中亦發現 LSTM 輸出須經過全連接層轉換為Q,K,V，因此本研究改進之核心部分在於作者提出改造後的前饋層(Boom Layer)。

✧ Boom Layer

為了減少運算量，基於Transformer前饋層改造後的 Boom Layer將原本前饋層透過激活函數GeLU轉換成 N 倍向量，分成 N 等分後加總將維度轉換回原維度。

$$v \in \mathbb{R}^H \longrightarrow u \in \mathbb{R}^{N \times H} \longrightarrow w \in \mathbb{R}^H$$

<< 維度轉換示意圖

五、RoBERTa: A Robustly Optimized BERT Pretraining Approach

✧ RoBERTa 之主要改善

將 BERT 之訓練過程做部分修改，包含增大 batch size，使用動態遮罩 (Dynamic Masking)，與應用更多語料進行訓練。修改 BERT 之主因在於其自身訓練不足，因此 RoBERTa 使用多種不同的最佳化方法改善預訓練性能。

✧ 靜態遮罩與動態遮罩

原始 BERT 於資料預處理 (Data Preprocessing) 即使用遮罩。

然而，RoBERTa 為了避免不同 epochs 使用到相同遮罩，提出 10 種不同的遮罩，共訓練 40 epochs，每筆資料用到 4 次相同遮罩，以於訓練過程增加更多變異。

✧ 動態遮罩

動態遮罩於模型訓練前動態生成遮罩模式 (Masking Pattern)，實驗顯示使用動態遮罩的訓練效能比靜態遮罩更好。

✧ 更大 batch size

BERT 原始論文中使用的 batch size 為 256，RoBERTa 使用 2K 及 8K 減少運算的 steps。此外，更大之 batch-size 更容易進行平行化運算。實驗過程發現，提高 batch size 至 8K 能減低 Masked Language Model 的困惑度 (Perplexity)。

三、資料集來源

資料集蒐集目的與預期結果旨在透過多維度的公開資料微調及訓練模型，提升模型在跨資料集和其他情感分析的泛化能力。此外，透過二分類的情感分析資料集提升模型對於強烈情緒語句的文本分類能力。

- SST-5 Fine-grained classification

  本資料集為 SST-5 Fine-grained classification with Rotten Tomatoes Movie Reviews 電影評論情感分類集，情感標籤包含5種情感等級:

Label	Comments
0	Negative
1	Somewhat negative
2	Neutral
3	Somewhat positive
4	Positive

  使用該資料集訓練不僅提供多維度情感分析，細粒度的情感標籤亦在訓練過中降低模型可能衍伸之情感等級模糊問題。

- SST2: Stanford - IMDb Dataset of 50K Movie Reviews

本資料集包含五萬餘則 IMDb 電影評論情感分類集，情感標籤有正面及負面2極情感維度。訓練集及測試集的文本序列數量分別各為25,000餘則。

四、研究方法及步驟

- **以 BERT 對原始文本進行預訓練針對特定下游任務進行微調**

BERT 之運作過程主要基於未標註或只有少量標註之文本數據進行微調以適用於特定下游任務。

運作過程包括以下三個主要步驟:

1. 準備原始文本數據:

文本數據包含未標註或少量標註之文本，透過數據清理將文本中空白標題的範例與 XML/html 之無效標籤去除，同時將超出BERT模型中預設序列長度的範本去除，並以將小於序列長度之向量補0，以符合預訓練文本固定序列長度讀入。

2. 將原始文本轉換成BERT相容之輸入格式:

文本進行預處理過程中對句子開頭向量位置加入分類符 [CLS]，並以[SEP]區分前一句與下一句。再以中文 BERT 對文本進行斷詞，如圖5。

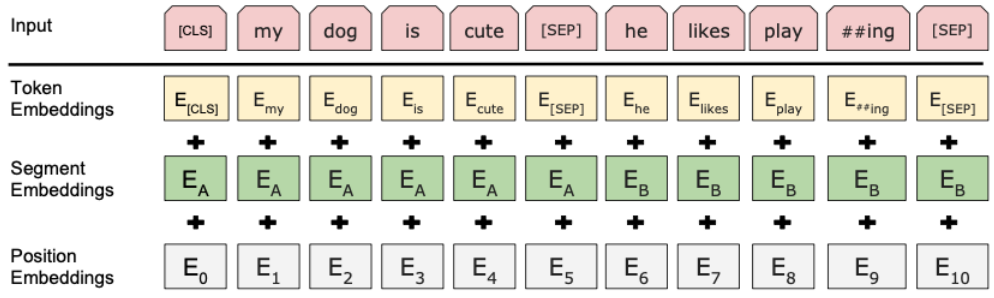


圖 5,BERT 成對句子編碼示意圖, 來源: BERT 論文

3. 於BERT頂層加入新的Layer進行微調使其適用於特定下游任務:

對BERT模型進行微調包含利用下游任務的目標函式從頭訓練分類器並微調BERT參數，以訓練完的BERT加上線性分類器最大化當前下游任務的目標，如圖6。

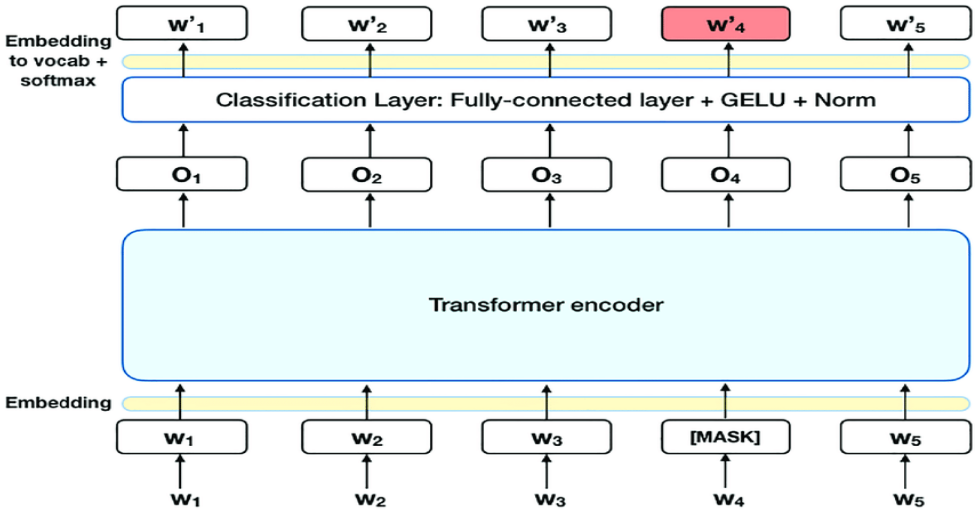


圖 6,BERT分類層示意圖, 來源: Faiza Khattak

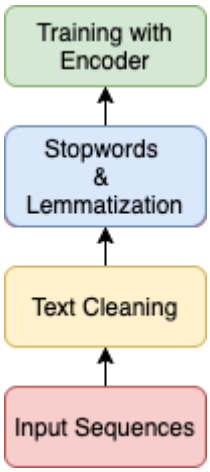
透過遷移學習，新增的分類器大多參數均來自已經預訓練的BERT模型，因此實際上需要重新訓練之參數量較少。微調的過程則依不同下游任務需求加入不同線性分類器。

◆ 實驗步驟

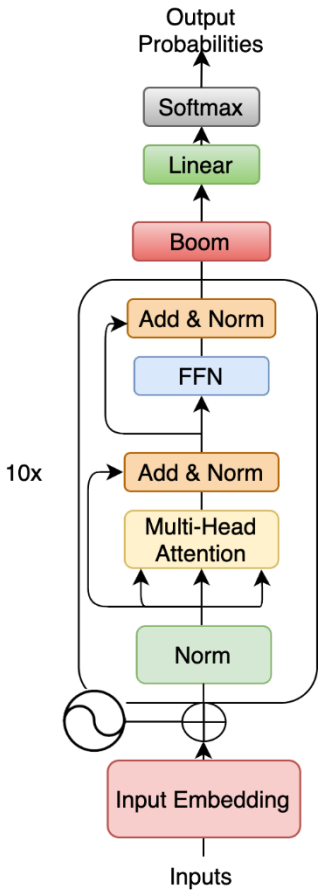
在基於Transformer之模型中挑選對情感分類資料集表現最佳的模型做為主要分類器與進行後續修改

實驗模型主要以基於Transformer的四種模型進行比較，此使用 BERT, RoBERTa, XLNet, DistilBERT 對情感分類數據集做分類。

挑選效能最佳之 RoBERTa 做為主要模型，並以Boom Layer改造前饋層做高維度轉換，提取更多的文本特徵與降低整個矩陣運算量。



對文本數據進行預處理



模型架構圖: EDM-RoBERTa

Fine-tuning Transformer-based Models with IMDb Dataset

	Epoch	Accuracy	train loss	valid loss	error rates
BERT$_{LARGE}$	6	92.6	0.35	0.55	0.29
RoBERTa$_{LARGE}$	6	93.17	0.22	0.53	0.26
XLNet	6	89.53	0.28	0.69	0.37
DistilBERT	6	86.48	0.32	0.74	0.35
EDM-RoBERTa	6	94.76	0.27	0.49	0.2

<<以 EDM-RoBERTa 與其他實驗樣本對 **SST-2: IMDb Dataset** 進行實驗

Fine-tuning Transformer-based Models with Rotten Tomatoes Dataset

	Epoch	Accuracy	train loss	valid loss	error rates
$BERT_{LARGE}$	5	66.21	0.64	0.68	0.3
$RoBERTa_{LARGE}$	5	68.91	0.67	0.7	0.29
XLNet	5	62.83	0.73	0.79	0.38
DistilBERT	5	54.65	0.8	0.77	0.44
EDM-RoBERTa	5	76.18	0.64	0.62	0.26

<<以EDM-RoBERTa與其他實驗樣本對SST-5: Rotten Tomatoes Dataset 進行實驗

EDM-RoBERTa (Enhance the Dependency Mechanism of RoBERTa)

bsz	steps	lr	ppl	SST-2	SST-5
256	1M	1.00E-05	3.83	92.6	74.57
2K	**125K**	**2.00E-04**	**3.61**	**94.76**	**76.18**
8K	31K	1.00E-03	3.72	92.1	74.31

<< EDM-RoBERTa 之訓練細節與效能比較

實驗過程以 SST-2 與 SST-5 資料集驗證模型效能，實驗環境如下方表格，訓練期間以 Wiki-Text103 與 CC-NEWs對 EDM-RoBERTa 進行預訓練，後續以模型對 SST-3 及 SST-5 進行監督式微調完成模型訓練。

- **實驗環境**

Google Colaboratory Pro	**GPU:** NVIDIA Tesla V100-SXM2-16GB
	CPU: Intel Xeon(R) @2.00GHz **OS:** Ubuntu 18.04.5 LTS **RAM:** 32GB
MacBook Pro (16-inch Late 2019)	**GPU:** AMD Radeon Pro 5600M-8GB-HBM2
	CPU: Intel Core i9-9980HK @2.4GHz **Dual Boot OS:** Ubuntu 18.04.5 LTS with macOS Catalina 10.15.7 (19H2) **RAM:** 64GB
MacBook Pro (13-inch, M1, 2020)	**GPU & CPU:** Apple M1 Chip with 8-core CPU, 8-core GPU **NPU:** Apple M1 Chip with 16-core Neural Engine **Environments:** CreateML, Tensorflow-mac **RAM:** 16GB

五、討論

研究與實驗結果顯示，透過本研究提出的 EDM-RoBERTa 模型對多維度情感資料集分析，得到比原始基於 Transformer 之模型(包含 BERT、RoBERTa、DistilBERT、XLNet)詞義分析輸出更精準的預測結果，證明 EDM-RoBERTa 有助於大幅改善 Transformer 注意力機制導致的編解碼過程無序缺失及其弱於捕獲文本中的短期依賴問題。本研究所獲得之成果將輔助應用於情感分析、社交網路分析等其他自然語言情感分析任務。

六、參考文獻

1. Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv: 1810.04805v2, 2019.
2. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jacob Uszkoreit, Llion Jones, Aidan N.Gomez, Lukasz Kaiser and Illia Polosukhin. Attention Is All You Need. arXiv: 1706.03762v5, 2017.
3. Yoav Goldberg. Assessing BERT's Syntactic Abilities. arXiv: 1901.05287v1, 2019.
4. Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng and Christopher Potts. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank, 2013.
5. Stephen Merity. Single Headed Attention RNN: Stop Thinking With Your Head. arXiv: 1911.11423v2, 2019.
6. Ilya Sutskever, Orial Vinyals and Quo V.Le. Sequence to Sequence Learning with Neural Networks. arXiv: 1409.3215v3, 2014.
7. Nikita Kitaev, Lukasz Kaiser and Anselm Levskaya. Reformer: The Efficient Transformer. arXiv: 2001.04451v2, 2020.
8. Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V.Le and Mohammad Norozui. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. arXiv: 1609.08144v2, 2016.
9. Stephen Merity, Caiming Xiong, James Bradbury and Richard Socher. Pointer Sentinel Mixture Models. arXiv: 1609.07843v1, 2016.
10. Lajanugen Logeswaran and Honglak Lee. An Efficient Framework for Learning Sentence Representation. arXiv: 1803.02893v1, 2018.