Enhancing the Dependency Mechanism of RoBERTa

專題組員： 李昱廷、郭為軒、曹仲辰、吳岳霖、林裕峰
指導老師：張炎清 教授
專題編號：PRJ2020-002

摘要

Our proposed model, EDM-RoBERTa uses SHA-RNN to improve the Multi-headed Attention mechanism in the encoder of Transformer model. Compared with the original Transformer and SHA-RNN models, the reorganization of Boom Layer and RoBERTa can meet both long and short text sequence input, and keep the original long-term dependency of Transformer. In the calculation process, it can also reduce the amount of calculation, improving the accuracy and text classification performance. The research source are Transformer-based models (BERT, RoBERTa, XLNet, DistilBERT), SHA-RNN, and the self-attention mechanism in Transformer as the main axis. The Boom Layer of SHA-RNN is transformed to realize the attention mechanism for high-dimensional vectors convertion, improve the original multi-head attention mechanism in the encoder of Transformer.

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

1. 簡介

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。

2. 專題進行方式

從二年級下學期選擇專題組員與指導老師後，於研究期間學習自然語言處理相關課程。在三年級開始了解分類器技術，開始確認研究主題主要應用領域的方向，後續亦完成並通過科技部計畫。專題進行方式為每週與專題指導老師 meeting，每次 meeting 時依照專案時程規劃提交完成進度並檢討開發上的問題。Meetinge 過程中除了小組方式進行，也由組員分享自學技術並共同討論。

2.1 人員配置與職責

人 員	工作與職責	人 員	工作與職責	人 員	工作與職責
李昱廷	負責組員間溝通聯絡 督促工作進行 系統設計 研究計畫規劃 論文投稿 文件製作、程式撰寫	郭為軒	文案校對 專題海報製作 文件製作 專題海報講解	曹仲辰	專題海報製作 文案校對 專題海報講解

人 員	工作與職責	人 員	工作與職責	人 員	工作與職責
林裕峰	專題海報講解 文案校對 文件製作	吳岳霖	文案校對 專題海報講解 專題海報製作 文案校對		

2.2 時程規劃

時間	工作項目
01/31~03/23 學習及熟悉技術工具	1、學習機器學習分類器與實作(Tensorflow、Python、Pytorch)。 2、了解分類器功能之技術。
03/24~04/30 確定專題目標階段	1、確認應用領域。 2、尋找 Benchmark Dataset。 3、決定專題題目。
05/01~06/20 系統分析與設計階段	1、確定論文改進方向。 2、規劃系統作業流程圖。 3、期末口頭與書面報告之準備。
06/21~09/21 製作階段	1、建立研究內容架構。 2、分類功能之設計。 3、寫系統程式。 4、分類功能改進。
09/22~10/31 評估階段	1、系統各個階段進行測試。 2、系統各個階段做適當的調整。 3、Transformer 與 SHA-RNN 之整合。
10/31~ 系統維護與展示階段	1、文件編修。 2、系統調整維護。 3、準備畢業報告。

3.研究動機與研究問題

3.1 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) 發展預訓練模型 (Pre-trained Model)以輔助人機之間的溝通理解及後續的自然語言生成 (NLG, Natural Language Generation)， 因此語義特徵的提取能力尤為重要。

3.2 研究問題

本研究使用重組後的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。訓練過程即判斷兩句話是否有相同含義，並捕捉句子間的關係。透過 SHA-RNN 降低運算量，實作單頭注意力解決 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。

4. 資料集來源

資料集蒐集的目的與預期結果在於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。

(1) SST-5 Fine-grained classification

本資料集為 SST-5 Fine-grained classification 電影評論情感分類集，其中情感標籤包含五種等級:

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。

(2) 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兩極情感維度。訓練集及測試集的文本序列數量分別各為 25,000 餘則。

5. 研究方法與實驗步驟

5.1 研究方法

以 BERT 對原始文本進行預訓練針對不同下游任務進行微調

BERT的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解決新的下游任務。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去除，同時將超出BERT模型中預設序列長度的範本去除，並以 0 將將小於序列長度的向量補 0，以符合預訓練文本讀入。

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。

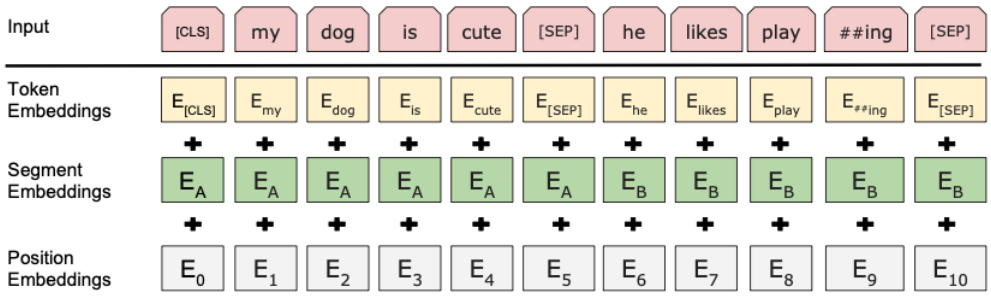


圖 5,BERT 成對句子編碼示意圖, 來源: BERT 論文

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。

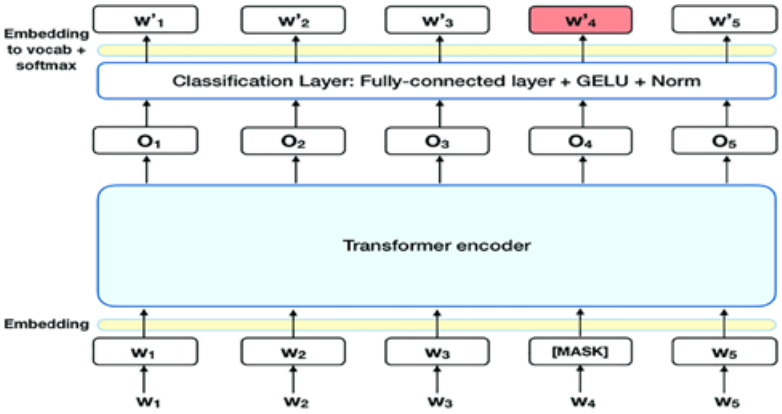


圖 6,BERT 分類層示意圖, 來源: Faiza Khattak

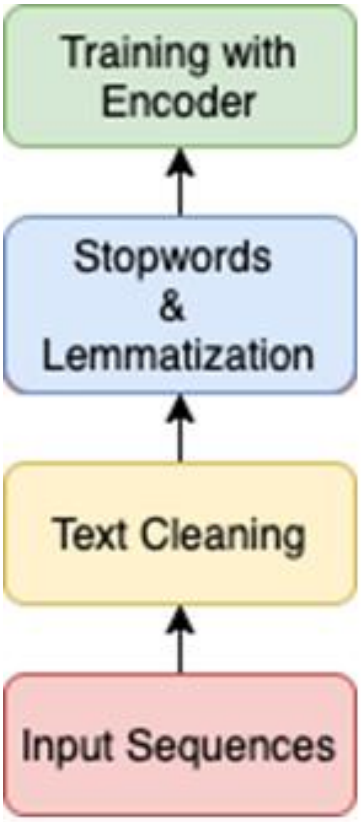
透過遷移學習，新增的分類器大多數的參數都來自已經預訓練的 BERT，因此實際上需要從頭訓練的參數量較少。微調的過程依不同下游任務加入不同之線性分類器。

5.2 實驗步驟

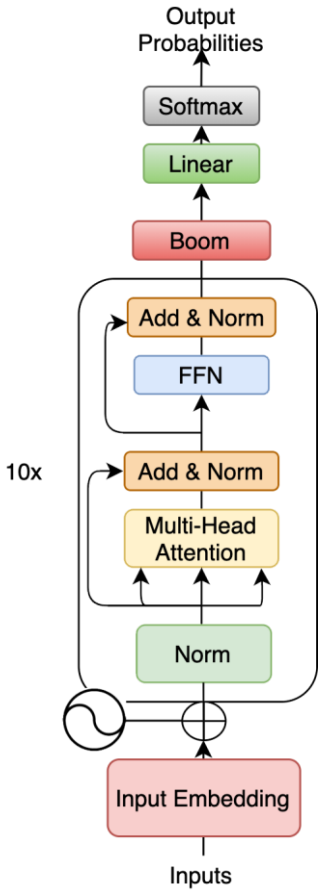
在基於Transformer之模型中挑選對情感分類資料集表現最佳的模型做為主要分類器及修改模型

實驗模型主要以基於Transformer的四種模型進行比較，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 做為主要的模型修改，以Boom Layer做高維度轉換提取更多的文本特徵降低整個矩陣運算量。



對文本數據進行預處理



模型架構圖

實驗過程以 SST-2 與 SST-5 資料集驗證模型效能，實驗環境為 Google Colaboratory (GPU: Nvidia V100, 16GB) 與 Intel Core i9-9980HK with AMD Radeon Pro 5600M 8GB HBM2，以 Wiki-Text 與 CC-NEWs 語料對 EDM-RoBERTa 進行預訓練，後續以本模型對 SST-3 及 SST-5 進行監督式微調完成模型訓練。

6. 實驗結果

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

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

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

7. 評估與展望

研究結果顯示，EDM-RoBERTa 得出比原始 Transformer 詞義分析輸出更精準的預測結

果，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。

8. 結語

終於到寫結語的時候了，想想還真不容易。在實驗室待了一年半載，研究方向從遺傳演算法換到自然語言處理。論文從一開始根本看不懂到現在準備專題發表，回首才驚覺自己已然走完這漫漫長路即將畢業，看著學長姐們的背影，今年也終於輪到我們了。
時光飛逝，歲月如梭，痛苦的時間確實比較難過，現在的感觸難以用隻字片語形容，那不如就寫到這裡吧，期許我們都有美好的未來，分道揚鑣，各自閃耀，乾杯，再見!

9. 銘謝

結語寫完了，感謝放在這裡。
特別感謝炎清老師接受我們的研究方向轉換，一路上給予我們的支持、指教與協助。
也感謝老師對我們鮮少的鴿子行為予以容忍，特別特別感謝!
最後，祝老師身體健康、工作順利、聖誕快樂、新年快樂，再會!

10. 參考文獻

Data Source	References Styles
Paper	Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv: 1810.04805v2, 2019.
Paper	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.
	Yoav Goldberg. Assessing BERT's Syntactic Abilities. arXiv: 1901.05287v1, 2019.
	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.
Paper	Stephen Merity. Single Headed Attention RNN: Stop Thinking With Your Head. arXiv: 1911.11423v2, 2019.
Paper	Ilya Sutskever, Orial Vinyals and Quo V.Le. Sequence to Sequence Learning with Neural Networks. arXiv: 1409.3215v3, 2014.
	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.
	Stephen Merity, Caiming Xiong, James Bradbury and Richard Socher. Pointer Sentinel Mixture Models. arXiv: 1609.07843v1, 2016.
Paper	Lajanugen Logeswaran and Honglak Lee. An Efficient Framework for Learning Sentence Representation. arXiv: 1803.02893v1, 2018.
	Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyou Zhou, Wenhu Chen, Yu-Xiang Wang, Xifeng Yan arXiv: 1907.00235v3, 2020.