運用混合注意力生成對抗網路於科學論文引用意圖分類
Generative Adversarial Networks based on Mixed-Attentions for Citation Intent Classification in Scientific Publications

王昱翔 Yuh-Shyang Wang, 陳昭沂 Chao-Yi Chen, 李龍豪 Lung-Hao Lee
國立中央大學電機工程學系
Department of Electrical Engineering
National Central University
{107521135, 107501543}@ncu.edu.tw, lhlee@ee.ncu.edu.tw

摘要

我們提出一個基於混合力的生成對抗網路模型 (簡稱 maGAN)，應用於科學論文引用分類任務。我們先選擇合適的領域訓練資料，透過提出的混合注意力機制，以及生成對抗網路架構，先預訓練語言模型，然後微調至多元分類任務。我們用 SciCite 資料集比較模型效能，由實驗結果得知 maGAN 模型得到最好的 Macro-F1 0.8532。

Abstract

We propose the mixed-attention-based Generative Adversarial Network (named maGAN), and apply it for citation intent classification in scientific publication. We select domain-specific training data, propose a mixed attention mechanism, and employ generative adversarial network architecture for pre-training language model and fine-tuning to the downstream multi-class classification task. Experiments were conducted on the SciCite datasets to compare model performance. Our proposed maGAN model achieved the best Macro-F1 of 0.8532.

關鍵字：注意力機制、預訓練語言模型、引用意圖、科學論文

Keywords: attentions, pretrained language models, citation intents, scientific publications.

1 緒論

近年來科學論文出版量大幅成長，透過自然語言處理的方法，分析探討這些資料變得相當重要，目前以深度神經網路為主要方向，但這需要大量標記的資料，大規模數據通常是交由群眾外包 (crowdsourcing) 的方式獲得，但要對科學文本進行手動標註，標記人員須要具備專業知識，導致蒐集標記成本極高。ELMo (Peters et al., 2018)、GPT (Radford et al., 2018) 和 BERT (Devlin et al., 2019) 等這些語言模型在大型語料庫上進行無監督預訓練，為許多的自然語言處理任務帶來顯著的提升。以 BERT 為例，BERT 訓練時不需要額外的標記資料，只以訓練文本中挑選兩句辨識是否為前後文。這類無監督預訓練的方法對於像科學、醫學這些無法獲得大量標記的領域變得相當重要。

SciCite (Cohan et al., 2019) 是艾倫人工智慧研究所提供的資料集，來源為 Semantic Scholar 語料庫 (Ammar et al., 2018)，標註關於科學論文的引用意圖分類，對於論文的引用可以分為背景知識、方法和結果比較，每筆論文引用屬於三種類別其中之一。現有的方法以使用 SciBERT (Beltagy et al., 2020) 或 BERT 進行微調為主，並以 Macro-F1 作為評分標準。

本研究使用 SciCite 作為評估預訓練模型效果的測試資料。我們透過使用科學文本 S2ORC 資料集 (Lo et al., 2020) 訓練出的基於混合注意力(mixed-attention, ma)的生成對抗網路

(Generative Adversarial Network, GAN)，模型簡稱 maGAN，在 SciCite 資料集上獲得的 Macro-F1 為 0.8532，比 BERT 的 0.844 和 SciBERT 的 0.8499 更高。

本文章節如下，第二章探討相關研究，第三章敘述我們提出的 maGAN 模型，第四章為實驗結果與分析，最後是結論。

2 相關研究

預訓練語言模型 (pretrained language models) 是透過大規模未標記語料訓練模型，接著在下游任務上微調模型。最初的語言模型以學習單獨的單詞表示為主，例如：Word2Vec (Mikolov et al., 2013) 以及 GloVe (Pennington et al., 2014)。後來藉由 LSTM (Hochreiter and Schmidhuber, 1997) 為基底，建立了 CoVe (McCann et al., 2017) 和 ELMo (Peters et al., 2018) 這類包含上下文資訊的單詞表示方式。近年來，基於多頭注意力(multi-head attention) 的 Transformer (Vaswani et al., 2017) 架構，在許多自然語言處理任務中，表現得比 LSTM 來得更好。GPT 採用 Transformer 作為主架構，加入生成訓練的概念，在下游任務有不錯的提升。Google 於 2019 年提出的 BERT 模型架構基於多層雙向的 Transformer，訓練過程中不採用傳統由左到右的語言建模方式，而是在兩個任務上進行訓練：遮罩語言模型 (Masked Language Model, MLM) 以及下一句預測(Next Sentence Prediction, NSP)。MLM 是將句子中的片段遮起來，預測遮蔽處應填入的字詞，NSP 則是判斷兩個句子當中，第二句是否為第一句在原始文本中的下一句。由於 NSP 需要輸入兩段文字作為訓練資料，需要的訓練資源較高，ELECTRA (Clark et al., 2020) 提出元素替換檢測 (replaced token detection)，透過模型將輸入文句的部分元素做替換，再判斷文句是否經過替換，降低訓練成本，也提升訓練效果。後續研究也有對注意力進行改良，ConvBERT (Jiang et al., 2020) 則是基於 ELECTRA 架構，加入卷積的概念收集區間訊息。

語言模型在下游任務的表現，與其所使用的訓練文本涵蓋領域高度相關，大部分的 BERT 相關模型，都是通用語料庫如 Wikipedia、Common Crawl 中訓練，因此在特定領域的任務表現相對較差。目前有許多特定領域的預訓練模型，比方說在生物醫學領域的 BioBERT (Lee et al., 2020)、使用臨床診斷書和出院報告的 Clinical-BERT (Alsentzer et al., 2020)，以及使用科學論文的 SciBERT (Beltagy et al., 2020) 等等。這些預訓練模型，都在與該領域相關的任務中獲得良好的表現。

本研究採用的預訓練資料為 SciBERT 團隊於 2020 年提出的開放研究語料庫 (The Semantic Scholar Open Research Corpus, S2ORC) (Lo et al., 2020)，該語料庫的建立是使用 Semantic Scholar 的文獻語料庫 (Ammar et al., 2020)，包含來自 MAG (Shen et al., 2018)、arXiv 及 PubMed 等約 811 萬篇來自各領域的英文論文，其中占比最多的三個領域為藥學、生物學及化學。

3 研究方法

3.1 注意力機制

Bahdanau et al. (2014) 提出注意力機制，為解決使用 Seq2Seq 模型時，編碼輸入越長的句子，越前面的資訊越容易丟失的問題，注意力機制將上下文以及位置資訊也一併傳遞下去，如此即便是模型末段，也能獲得在序列前端的重要資訊。

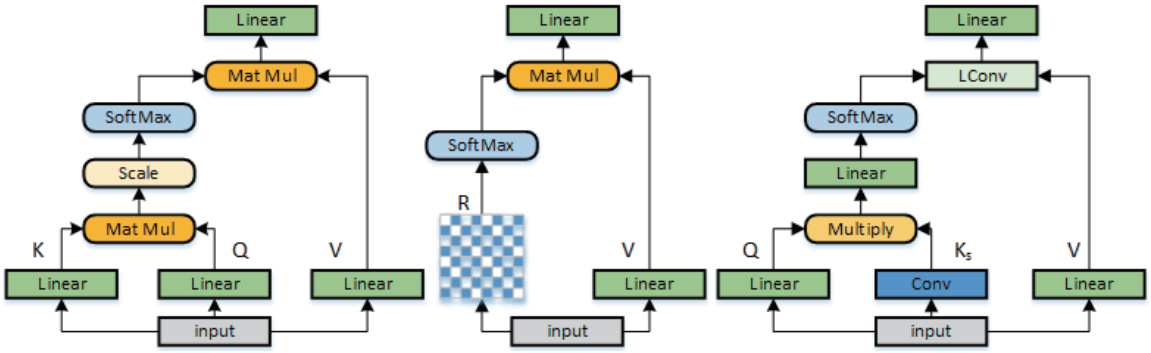
自注意力(self-attention) 如圖 1 (a)。計算過程如方程式 (1) 到 (4)，由輸入序列 A 產生三個矩陣 Q (Query), K (Key), V (Value)，先對 Q 跟 K 進行點積，相當於計算 Q 跟 K 的相似度，再將其降低 d_k 維度，並透過 softmax 轉換為機率分布，最後再乘上 V 獲得注意力權重。雖然自注意力具有學習非局部資訊的特色，但是仍有相當比例的注意力頭 (head) 是在學習局部依賴性，如果只採用部分注意力頭，反而會提升表現 (Kovaleva et al., 2019)。

$$Q = W_q A \quad (1)$$

$$K = W_k A \quad (2)$$

$$V = W_v A \quad (3)$$

$$SelfAttention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (4)$$



(a) self attention (b) random synthesized attention (c) span-based dynamic convolution

圖 1: 三種注意力機制

隨機合成注意力(random synthesized attention) (Tay et al., 2020) 是 synthesized attention 的一種如圖 1 (b)。計算過程如方程式 (5) 與 (6)，透過一個完全不受輸入 token 影響、隨機初始化的矩陣 R 進行訓練，矩陣 R 經過 softmax 得到機率分布後，再與由輸入序列產生的矩陣 V 相乘，獲得注意力權重。

$$V = W_v A \tag{5}$$

$$Random\ Synthesized\ Attention(R, V) = softmax(R)V \tag{6}$$

跨度動態卷積 (span-based dynamic convolution) 是由 ConvBERT (Jiang et al., 2020) 提出的注意力提取方式，如圖 1 (c) 所示。計算方式如方程式 (7) 與 (8)，該研究觀察到多頭注意力中，有些頭 (head) 只需要局部資料，便能完成注意力。因此，建立了局部依賴機制，藉此減少不必要運算量，但這種設計會在獲取全局資訊上表現較差。使用由 MobileNets 提出的深度分離卷積 (depthwise separable convolution) (Howard et al., 2017) 產生 K_s，再與 Q 做點積、降維、softmax 後，再與輸入序列產生的矩陣 V，一起經過輕量卷積 (light-weight convolution) (Wu et al., 2019)，就能得到跨度動態卷積的注意力權重，計算方式如方程式 (7) 與 (8)，其中 W 為 CNN 的卷積核。由於不同注意力機制各有優缺點，若能結合長處並填補短處，便能更有效的提取資訊。因此，我們提出的混合注意力 (mixed-attention) 是由三個區塊所組成，包含自注意力、隨機合成注意力及跨度動態卷積，透過多頭注意力 (multi-head attention)，將自注意力機制 (SA)、隨機合成注意力(RSA)及跨度動態卷積 (SDConv)，三者以不同頭數獲得的注意力權重連接起來，並獲得新的注意力權重，如方程式 (9)。

$$LConv(X, W, i) = \sum_{j=1}^{k} W_j \cdot X_{i+j-\left[\frac{k+1}{2}\right]} \tag{7}$$

$$SDConv() = LConv(V, softmax\left(W_f(Q \odot K_s)\right), i) \tag{8}$$

$$MixAttention() = Concate(SA(), RSA(), SDConv()) \tag{9}$$

3.2 模型架構

我們提出的混合注意力生成對抗網路 (mixed-attention-based Generative Adversarial Network, maGAN) 模型，使用大量科學論文 S2ORC 資料集 (Lo et al., 2020) 作為預訓練語料，以 ELECTRA 架構 (Clark et al., 2020) 為基底，並採用改良的混合注意力機制訓練語言模型，預訓練 (pre-training) 架構如圖 2 (a)，分為兩個主要部分：生成器 (Generator)、判別器 (Discriminator)。生成器的功能是將輸入句中的遮罩部分進行替換，產生新的句子，並作為判別器的輸入。而判別器則是判斷輸入句

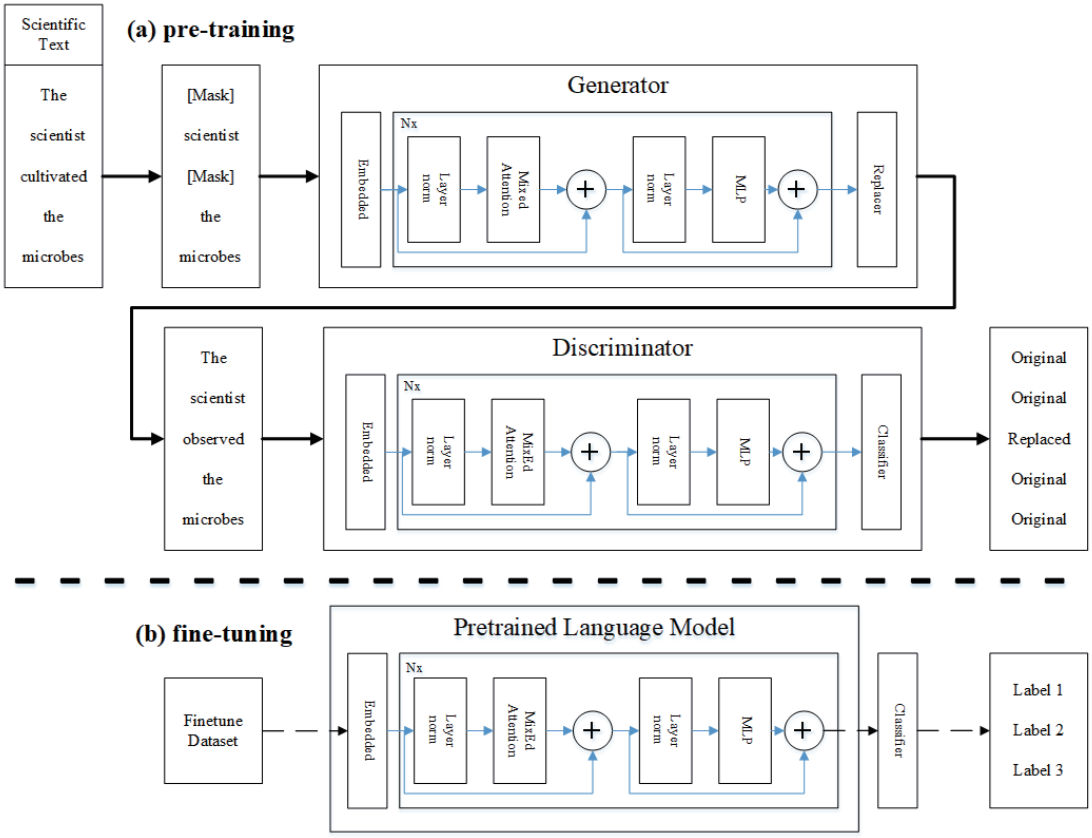


圖 2: 混和注意力生成對抗網路模型架構

中的字詞是否被生成器所替換。生成器和判別器架構相似都是透過其中的詞嵌入層 (embedding layer) 將輸入文字序列轉為固定詞向量，輸入編碼器 (transformer encoder)，藉由注意力機制獲取語意資訊，經過 N 層編碼器後，產生帶有語境的詞向量。

與 ELECTRA 架構不同的是，此處的注意力機制改用我們所提出的混合注意力 (mixed-attention)，而不是原來的多頭自注意力 (multi-head attention)。生成器提取資訊後，將資訊輸入到替換器 (replacer)，將文句進行替換，並輸出經替換的文句交由判別器；判別器完成資訊提取後，輸入到分類器中，判斷是否有元素被替換。

在圖 2 (b) 的微調 (fine-tuning) 階段，將下游任務的資料，藉由已完成訓練的判別器進行資訊提取，再連接符合任務需求的分類器進行訓練，微調過後的分類器，最後用來預測類別標籤 (label)。

4 模型評估

4.1 資料集

實驗資料來自 SciCite 資料集 (Cohan et al., 2019)，將科學論文的引用意圖分為三類：背景知識 (background information)、方法 (method) 與結果比較 (result comparison)，類別定義與範例如表 1 所示。其中訓練集含有 8,243 筆、發展集有 916 筆，而測試集包含 1,861 筆。

4.2 實驗設定

模型參數設定均與 ELECTRA-BASE 相同，只有 Batch size 礙於硬體需求降為 64。而模型中採用的混合注意力機制中，三種注意力 SA : RSA : SDConv 各自頭數以 3:3:1 的方式組合。

效能指標如同公開的效能評測排行榜 (leaderboard)，以 Macro-F1 作為主要的評分依據，先計算各個類別的 Precision 及 Recall，然後算其調和平均數 F1-score，再將各類別的 F1 平均，即可獲得 Macro-F1。

意圖類別	定義	範例
背景知識 Background information	引文提供有關問題、概念、方法或重要性的背景信息	Recent evidence suggests that co-occurring alexithymia may explain deficits [12]. Locally high-temperature melting regions can act as permanent termination sites [6-9]. One line of work is focused on changing the objective function (Mao et al., 2016).
方法 Method	使用方法、工具或數據集	Fold differences were calculated by a mathematical model described in [4]. We use Orthogonal Initialization (Saxe et al., 2014)
結果比較 Result comparison	論文的結果或發現與相關研究的比較	Weighted measurements were superior to T2-weighted contrast imaging which was in accordance with former studies [25-27] Similar results to our study were reported in the study of Lee et al (2010)

表 1: 引用意圖類別與範例

Model	Macro F1
BiLSTM-Attention+ELMo	82.6
Structural-scaffolds	84.0
SciBERT	84.99
maGAN (ours)	85.32

表 2: 模型效能評估結果

預測分類 \ 實際分類	背景知識	方法	結果比較
背景知識	884	93	25
方法	44	489	1
結果比較	69	23	233

表 3: 錯誤分析混肴矩陣

4.3 效能分析

實驗結果如表 2。其他模型效能取自公開的效能評測排行榜[1]，除了 SciBERT 外，其他模型為 Cohan et al. (2019) 提出資料集的同時所提出的方法，BiLSTM-Attention+ELMo 使用 ELMo 為詞嵌入並藉由 BiLSTM 提取注意力，而 Structural-scaffolds 則是基於前者同時訓練多個任務，我們的 maGAN 模型獲得了 Macro-F1 85.32 目前是最好的成績。

表 3 為錯誤分析的混肴矩陣，最容易辨識錯誤的狀況為「方法」類別辨別為「背景知識」(佔全部錯誤的 36.4%)，其次是將「背景知識」分辨為「結果比較」(佔 27%)。根據我們的觀察，有部分錯誤是因為缺乏全文資訊而導致誤判，若是額外將論文前後文加入微調訓練中，有機會正確判斷分類。

[1] https://paperswithcode.com/sota/citation-intent-classification-on-scicite

5 結論

本研究針對科學論文引用分類任務，提出一個基於混和注意力的生成對抗網路模型，透過選擇合適的領域訓練語料，提出混和注意力機制，透過生成對抗網路架構，先預訓練語言模型，然後微調至分類任務，在 SciCite 測試資料獲得 0.8532 的 Macro-F1，目前是表現最好的模型。

致謝

This work was partially supported by the Ministry of Science and Technology, Taiwan under the grant MOST 108-2218-E-008-017-MY3.

參考資料

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly Available Clinical BERT Embeddings. *arXiv Preprint*, https://arxiv.org/abs/1904.03323

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, et al. 2018. Construction of the literature graph in semantic scholar. In *Proceedings of NAACL-HLT'18, (Industry Papers)*. pages 84–91.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv Preprint*, https://arxiv.org/abs/1409.0473

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of EMNLP-IJCNLP'19,* pages 3615-3620.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv Preprint,* https://arxiv.org/abs/2003.10555

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Fiels Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. *arXiv Preprint*, https://arxiv.org/abs/1904.01608

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv Preprint*, https://arxiv.org/abs/1810.04805

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, *9*(8):1735-1780.

Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv Preprint*, https://arxiv.org/abs/1704.04861

Zihang Jiang, Weihao Yu, Daquan Zhou, Yunpeng Chen, Jiashi Feng, and Shuicheng Yan. 2020. ConvBERT: Improving BERT with Span-based Dynamic Convolution. *arXiv Preprint*, https://arxiv.org/abs/2008.02496

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. *arXiv Preprint,* https://arxiv.org/abs/1908.08593

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4):1234-1240.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of ACL'20*, pages 4969-4983.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. *arXiv Preprint,* https://arxiv.org/abs/1708.00107

Thomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS'13*, pages 3111-3119.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of EMNLP'14*, pages 1532-1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceddings of NAACL'18*, pages 2227-2237.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Zhihong Shen, Hao Ma, and Kuansan Wang. 2018. A Web-scale system for scientific knowledge exploration. In *Proceedings of ACL'18, System Demonstrations*, pages 87-92.

Yi Tay, Dara Bahri, Donald Metzler, Da-Cheng Juan, Zhe Zhao,and Che Zheng. 2021. Synthesizer: Rethinking Self-Attention for Transformer Models. In *Proceedings of ICML'21*, PMLR 139:10183-10192.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. *Attention is all you need.* In *Proceeding of NIPS'17*, pages 5998-6008.

Felix Wu, Angela Fan, Alexei Baevski, Yann N. Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. *arXiv Preprint*, https://arxiv.org/abs/1901.10430