Social Media Mining

[Final project] Paper citation network Graph classification

系級: 資工碩一
學號: 110526001
姓名: 鄭凱元

Abstract

學術期刊是一種經過學術團體、專業領域同儕評審的期刊，當一篇論文被收錄到期刊中，便有可能出現在學術平台上增加曝光率，而許多平台是一種評論文分類的系統，可利用統計文獻的方式來說明學術領域的研究趨勢與最新的技術發展，將任何人公開發表出來的學術研究進行篩選、分類，依據論文其目的、範圍領域、讀者群放置在正確的類別中供後人研讀，然而許多沒有收錄在期刊中的文獻往往都被收錄在各大學院校的文獻分類系統中，這些分類系統根據文獻的關鍵字、作者，甚至是更細部的資訊進行分類並回饋給使用者，但根據研究[1]顯示，市面上部分的論文分類系統常會把文獻分類至錯誤的學科，這可能誤導那些不熟悉該領域的人員學習，為了探討論文分類的實際問題與準確度，本研究列出了幾種情況下模型對於論文分類的優劣以及分析其輸出結果。

1 Introduction

隨著越來越多的資訊、理工、醫療相關研究的快速發展，帶給學術論文的類別更多的複雜型態。建立學術論文平台的目的，除了資加資料的透明度外，並且希望能促進研究及分享研究成果。跟google scholar的功能不同，一個好的期刊分類系統，他們能夠將蒐集到的文獻按照其內容分配到各個領域、主題，方便後人了解學科的結構和歷史發展。但是以Scopus、NSA ADS透過論文的journal來做分類的方法來說，此方法沒有辦法套用在還未發表在期刊上的論文，而對於剛進入學術界的研究人員來說，他們可能也沒辦法精確地定位自己的論文適合哪些期刊上。為了讓研究員更清楚地意識到自己的論文屬於哪一類型的期刊，藉由論文以及其參考文獻等資訊，本實驗設計一套BERT結合GCL的模型來實現論文分類系統，並與其他模型做評估比較，討論他們預測結果與分析。

2 Method

2.1 Experiment Dataset

Nature是世界上最早的科學期刊之一，在 nature平台上刊登了許多不同科學領域的文獻資訊，其中包含了關於氣候變遷、醫學、數據科學等相關研究，每篇文獻皆有的title、author、date、references和abstract，本實驗使用論文的title、author、date，利用字元(;)將這3個資訊合併成一個字串作為GNN中的node，而references作為edge輸入到模型中進行訓練。本實驗過程分為兩個階段，實驗一的階段本研究蒐集了Nature Machine Intelligence; Light: science & applications; Nature Computational Science; Scientific Data; Nature Biomedical Engineering等5篇不同類別的論文，並對各類別中重複出現的論文進行過濾篩選，共1200篇資料集；實驗二中本研究蒐集了Journal of Exposure Science & Environmental Epidemiology; Nature Medicine; Nature Energy; Nature Neuroscience; NPG Asia Materials; Nature Microbiology; Nature Geoscience等7篇相關領域的類別，並對各類別中重複出現的論文進行過濾篩選，共5312篇資料集(如表1)。

Name	Graph number	Avg. Node	Avg. Degree
NCI1	4110	29.8	1.08
PROTIENS	1113	39.06	1.86
My_data	**5312**	**17.68**	**2.85**

表1

Node example: yu, t., kim, d. y., zhang, h. & xia, y. n.；platinum concave nanocubes with high - index facets and their enhanced activity for oxygen reduction reaction.；(2011).

2.2 Embedding method

由於將單詞直接進行編碼，可能導致向量過長以及單詞間的關係過於稀疏，因此，將單詞轉成一組固定維度的Embedding方法，並藉由訓練模型來調整單詞在高維空間中的距離關係，可使單詞在該空間上獲得語意。Glove embedding是一種結合了LSA的矩陣分解方法以及word2vec的sliding window方法的embedding方法，採用字詞共現矩陣來統計字詞資訊，矩陣中的元素表示目標字詞及其sliding window內上下文共同出現的次數，當sliding window內的兩字詞距離愈遠則權重愈小，然而glove embedding model沒辦法對unseen word進行處理，另外考慮到memory不夠等問題，因此本研究通過tiny bert模型對文本先進行subword tokenize後再做embedding的任務。

2.3 Graph contrastive learning

對比學習是一種目前常使用的深度學習非監督學習方法，在圖神經網路中這項技術最早被應用在[2]的研究上作為pretraining method，首先對輸入的graph複製成兩份，分別隨機用node drop、attribute mask、edge remove、random walk增強graph的表示(本次實驗總共試了node dropping、random walk、Identity3種方法)，並將增強的graph pair輸入到GCN+MLP的hidden layer中，最後輸出兩個視圖的vector並透過contrastive loss來最大化一致性，在contrastive learning中正樣本就是一組 augmentation graph pair，然後這兩個graph來自同一篇論文，至於 nagative sample則是指graph pair的2個graph是來自不同的論文，因此在經過對比學習後，模型可以學到如何去區別出某篇論文它跟其他篇論文的不同處在哪。而在finetune的部分，GCL的原文作者取得pretraining中的GCN部分並作為graph的encoder，接著後面接了一個MLP網路作為下游任務的分類器(如圖1)。

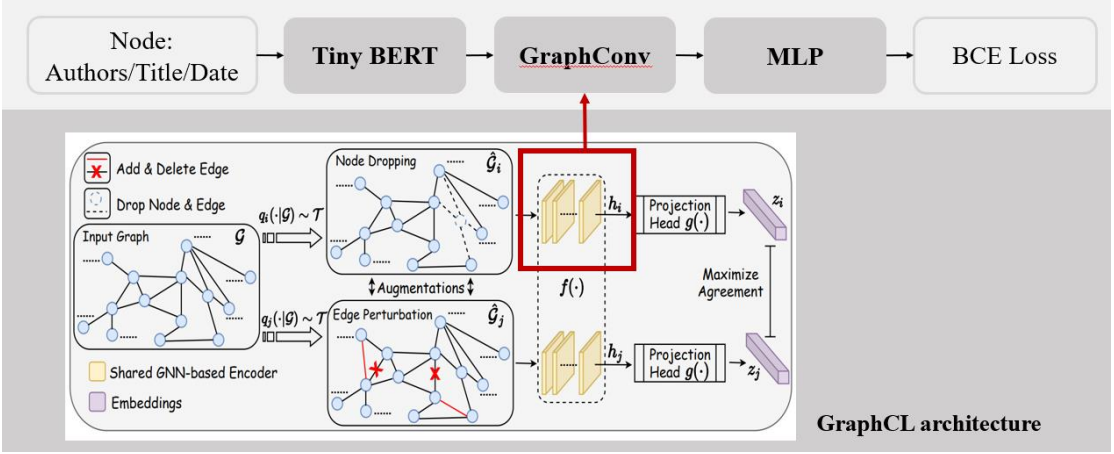


圖1

2.4 Model difference

本研究在embedding方法時使用Tiny BERT來進行node embedding外，並且在GraphCL的架構上做了部分修改，像是在GCN encoder的部分，因為本文所使用的paper citation network資料集沒有太複雜的結構(從central node到最遠的neighbor距離為2)，因此GCN的layer層數只用2(原文為3)；在finetune部分，為了保留pretraining時encoder對於每篇paper citation network的特徵表示，所以本文freeze GCN encoder的權重(原文中沒有使用)，只讓模型對下游任務的部分做學習與更新。

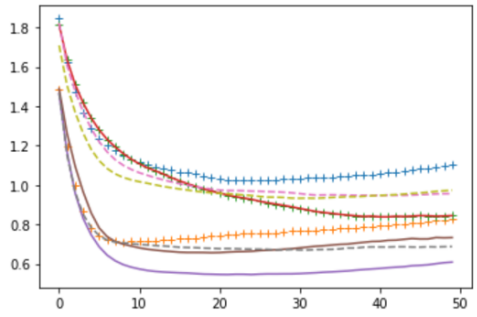
3 Result

3.1 Experiment1

實驗一蒐集了nature平台上的5篇學科上比較沒關係的journal分別為Nature Machine Intelligence; Light: science & applications; Nature Computational Science; Scientific Data; Nature Biomedical Engineering，透過上述實驗方法建立model並做5種類別的分類任務，並使用0.1%的label資料集進行finetune，得到的準確度為80.9%；另外，再將Tiny BERT encoding後的輸出結果再輸入到自訂的1層Dense layer做分類任務，準確率卻也達到81.5%。

3.2 Experiment2

實驗二蒐集了nature平台上的7篇學比較相關係的journal分別為Journal of Exposure Science & Environmental Epidemiology; Nature Medicine; Nature Energy; Nature Neuroscience; NPG Asia Materials; Nature Microbiology; Nature Geoscience，透過上述實驗方法建立model並做7種類別的分類任務，實驗中我總共測試了9種augmentation的組合，然後做finetune任務時，以NodeDrop+RW的pretraining為例，我會先拿NodeDrop的GCN來作為第一組finetune模型的GCN encoder，然後第二組finetune模型的GCN encoder會拿RW的GCN，實驗結果顯示，使用NodeDrop+RW pretraining模型中的RW GCN作為finetune的encoder，可以使finetune時的accuracy得到最高分(如圖2)。



Pretraining mode	Finetune Accuracy
NodeDrop+NodeDrop	0.837
NodeDrop+RW	0.831
RW+NodeDrop	**0.841**

紅色代表: 使用這個augmentation graph的GCN做finetune

圖2

3.3 Compare with model

原始論文中為了評估模型的performance，因此使用NCI1和PROTEINS資料集進行semi-supervised的任務， GCL model的相關資訊如3.4所述，本實驗在表2跟原始論文的模型比較NCI1和Proteins資料集的performance外，也跟原始論文比較在paper network上的表現，原始論文的模型使用one hot的方法來進行7篇論文的分類，而本實驗分別觀察了在Tiny BERT(author/title/date)、BERT(abstract)、BERT+GCL(with pretraining)、BERT+GCL(without pretraining)的表現；實驗結果如表3顯示，使用abstract的資訊話可以獲得更高的performance，然後在相似期刊的資料集中，會比較適合使用BERT+GNN-based的方法。

Model (Label Rate)	NCI1	Proteins
GraphCL (1%)	0.607	0.627
GraphCL (10%)	0.742	0.714
My Model (1%)	0.593	0.584
My Model(10%)	0.676	0.673

表2

Label Rate	Tiny BERT (date/title/author)	BERT (abstract)	GraphCL (one hot)	Bert + GCL	Bert + GCL (without Pretrain)
1%	0.690	0.663	0.300	0.626	0.628
10%	0.705	0.788	0.403	0.841	0.811

表3

4 Discussion

根據上述的結果，當分類的文獻相學科領域差距很大時，BERT分類的準確度相當於使用更複雜的BERT+GCL，而當使用類似的文獻進行分類時，由於輸入的文本過於相似，因此需要使用論文references的資訊來增加特徵才會有比較好的performance，另外node embedding的方法對於模型有明顯的影響，因為如果只使用one-hot方式訓練模型，node彼此間可能會過於稀疏導致很難學到graph的具體特徵，因此在GNN-based的模型中，embedding的方法對於最後的performance會有很大的影響；實驗還發現，在pretraining 的時候，雖然只更新分類器的網路(encoder的網路是freeze的狀態)，但依然可以獲得非常高的準確率，這代表 model的encoder已經學到辨識不同paper間的重要特徵，那這組 encoder 輸出的特徵有利於之後的分類器去做下游任務的學習。

從模型預測錯誤的結果可以發現，模型在預測特定的文獻時可能會預測成另一個特定的文獻，像是在energy方面的論文中，只要發生預測錯誤時都會被模型歸類到material方面的文獻上，又或是當文獻為neuroscience的時候，往往會被預測為medical的情形，這可能是因為我們給模型的特徵數量還不足以區分這兩個文獻間的差異，因此未來研究中如果能改善記憶體不夠的問題時，或許可以將目前的title、author等輸入資料，改成使用abstract相關的資料來進行模型訓練。

5 References

- Shu, F., Julien, C. A., Zhang, L., Qiu, J., Zhang, J., & Larivière, V. (2019). Comparing journal and paper level classifications of science. Journal of Informetrics, 13(1), 202-225.
- You, Y., Chen, T., Sui, Y., Chen, T., Wang, Z., & Shen, Y. (2020). Graph contrastive learning with augmentations. Advances in Neural Information Processing Systems, 33, 5812-5823.
- You, Y., Chen, T., Shen, Y., & Wang, Z. (2021, July). Graph contrastive learning automated. In *International Conference on Machine Learning* (pp. 12121-12132). PMLR.
- Chen, T., Kornblith, S., Swersky, K., Norouzi, M., & Hinton, G. E. (2020). Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, *33*, 22243-22255.

6 Others(colab pro+的GPU和大小)

一個帳號共可以開兩個notebook，一個notebook限用一張GPU:

```
1 !nvidia-smi
```

```
Sat May 21 13:02:13 2022
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 460.32.03    Driver Version: 460.32.03    CUDA Version: 11.2     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  Tesla P100-PCIE...  Off  | 00000000:00:04.0 Off |                    0 |
| N/A   47C    P0    36W / 250W |   1147MiB / 16280MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
+-----------------------------------------------------------------------------+
```