Social media mining
Final Project

Title : paper citation network Graph classification

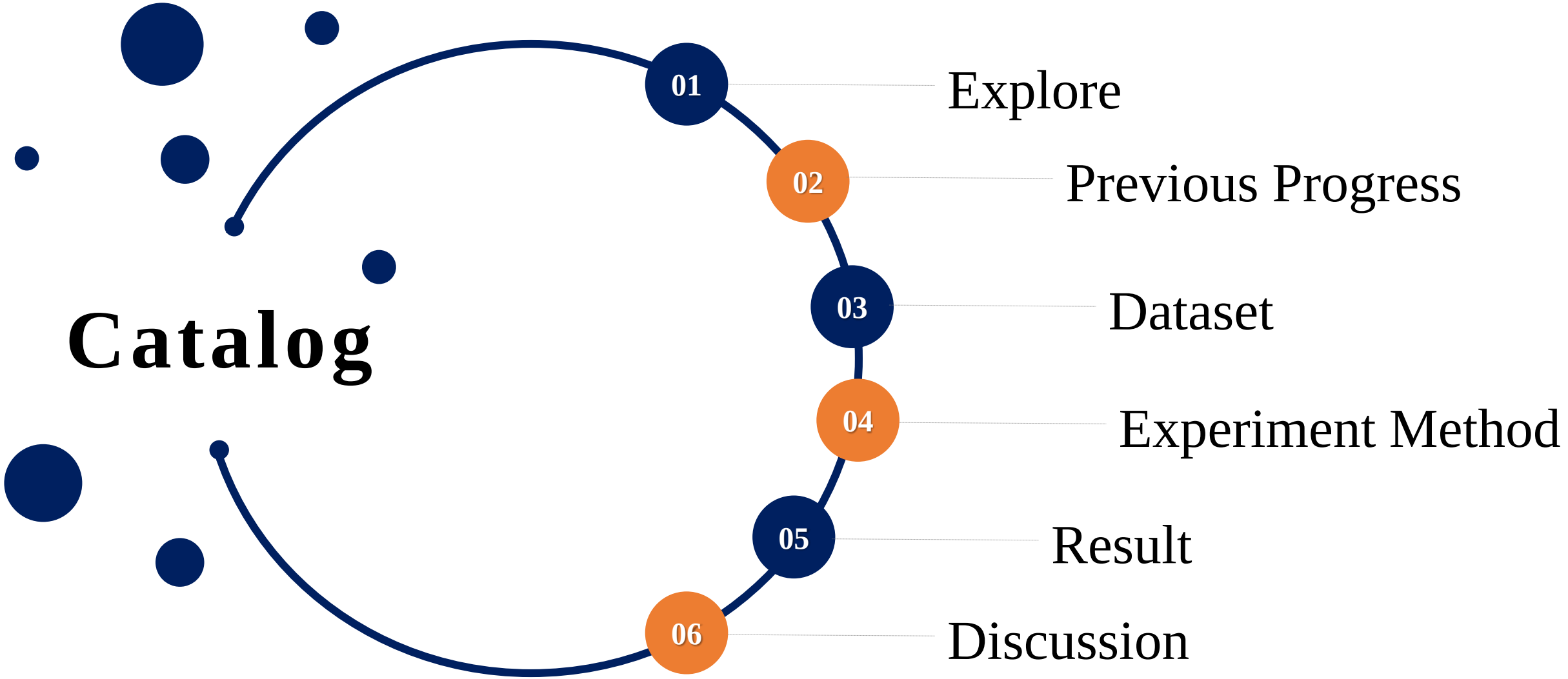
組別 : 第三組

系級 : 資工所碩一

學號 : 110526001

姓名 : 鄭凱元

Catalog


01
Explore
02
Previous Progress
03
Dataset
04
Experiment Method
05
Result
06
Discussion

Explore

1. Motivation: Which journal does the paper belong to?
2. BERT limitation on the paper classification
3. BERT + GCL performance
4. Discussion&Future improvement

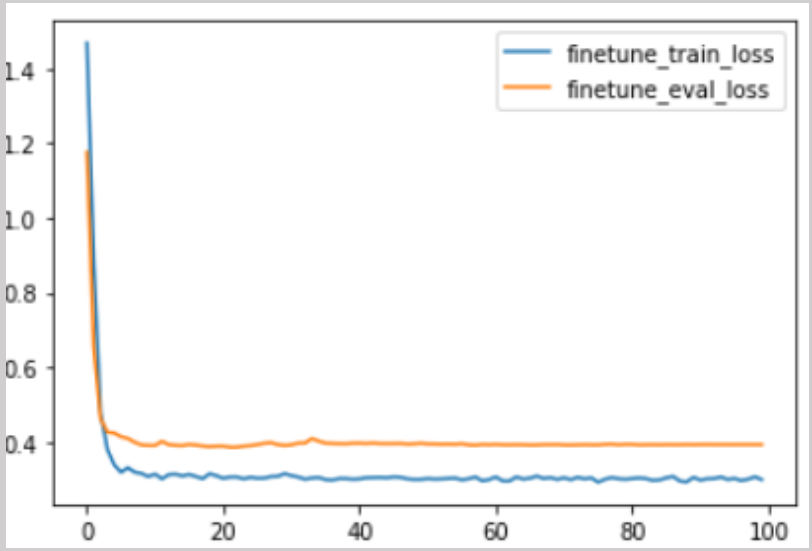
Previous Progress

5 journals:

- Nature Machine Intelligence
- Light: science & applications
- Nature Computational Science
- Scientific Data
- Nature Biomedical Engineering

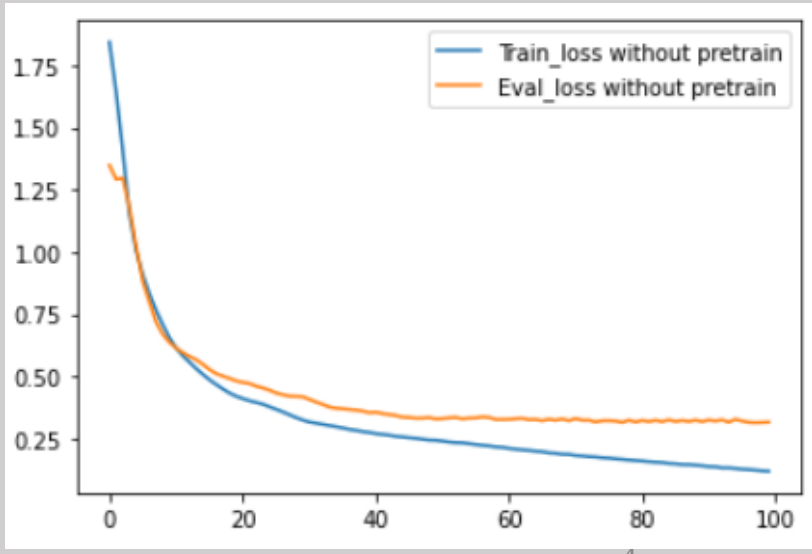
BERT

Accuracy: 0.815 (5 class)



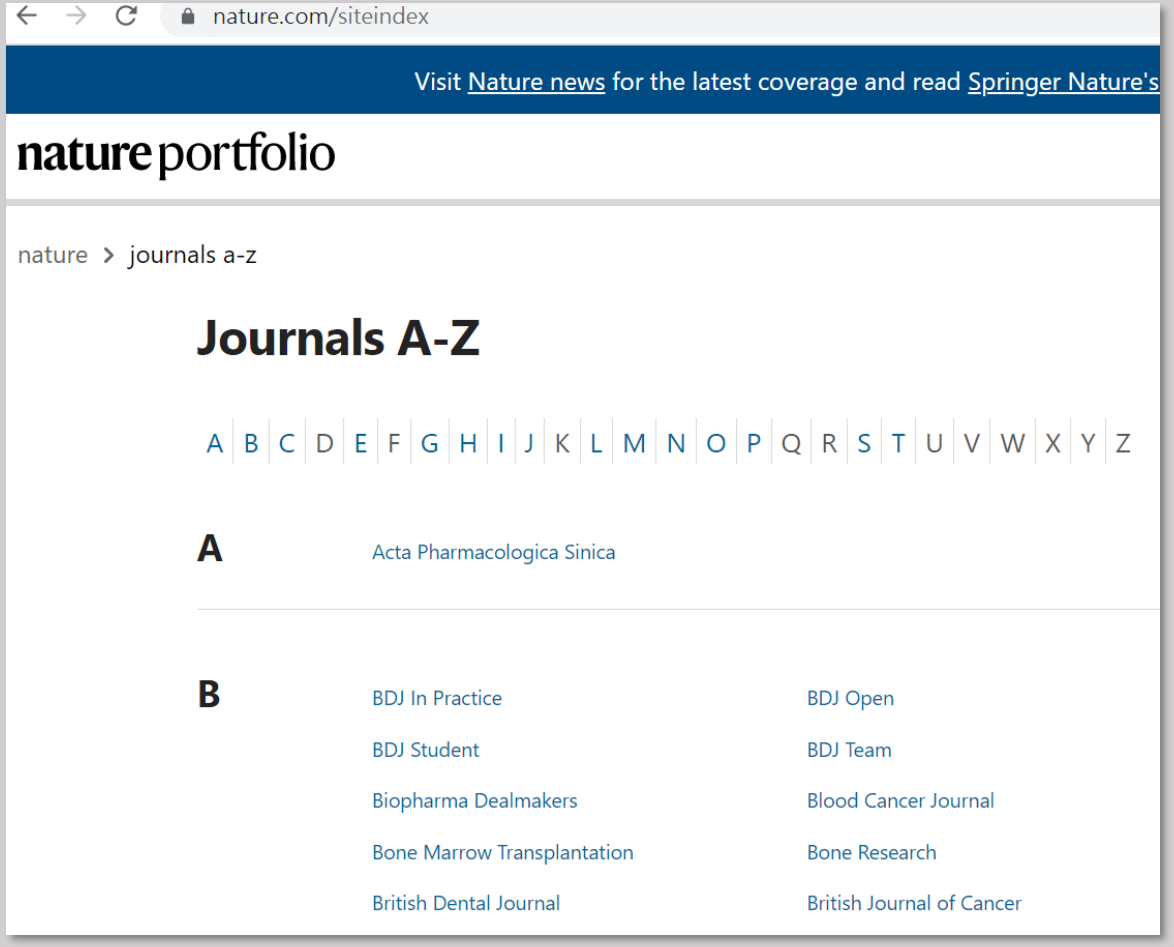
BERT+GCL

Accuracy: 0.809 (5 class)



4

Dataset



- Many Science, Medicine, Genetics and Biology papers are in “NATURE journal”.
- In the platform, each paper has its own category.
  1. Journal of Exposure Science & Environmental Epidemiology
  2. Nature Medicine
  3. Nature Energy
  4. Nature Neuroscience
  5. NPG Asia Materials
  6. Nature Microbiology
  7. Nature Geoscience

5

Dataset column

Subject (Label)

Date
Author
Title
} Input Data
(Concat)

url

Reference

Abstract

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**

Fine-tune Data (Label Rate 10%)

- Training data : 531 (10%)
- Evaluation data : 4781 (90%)

Fine-tune Data (Label Rate 1%)

- Training data : 53 (1%)
- Evaluation data : 5259 (99%)

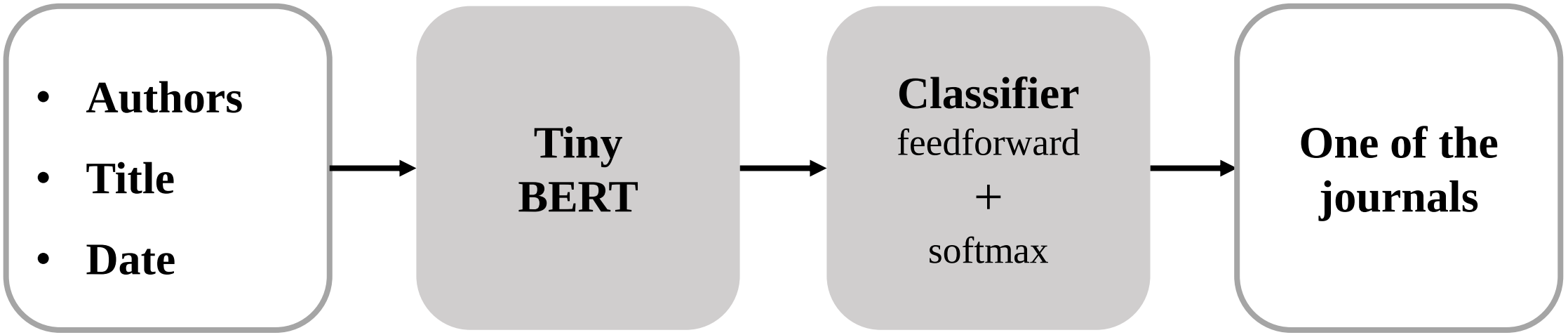
Select embedding methods

Some **proper nouns and abbreviations appear** in the dataset, making unseen word error for glove embedding.

Using BERT model, but will cause the problem of **out of memory**.

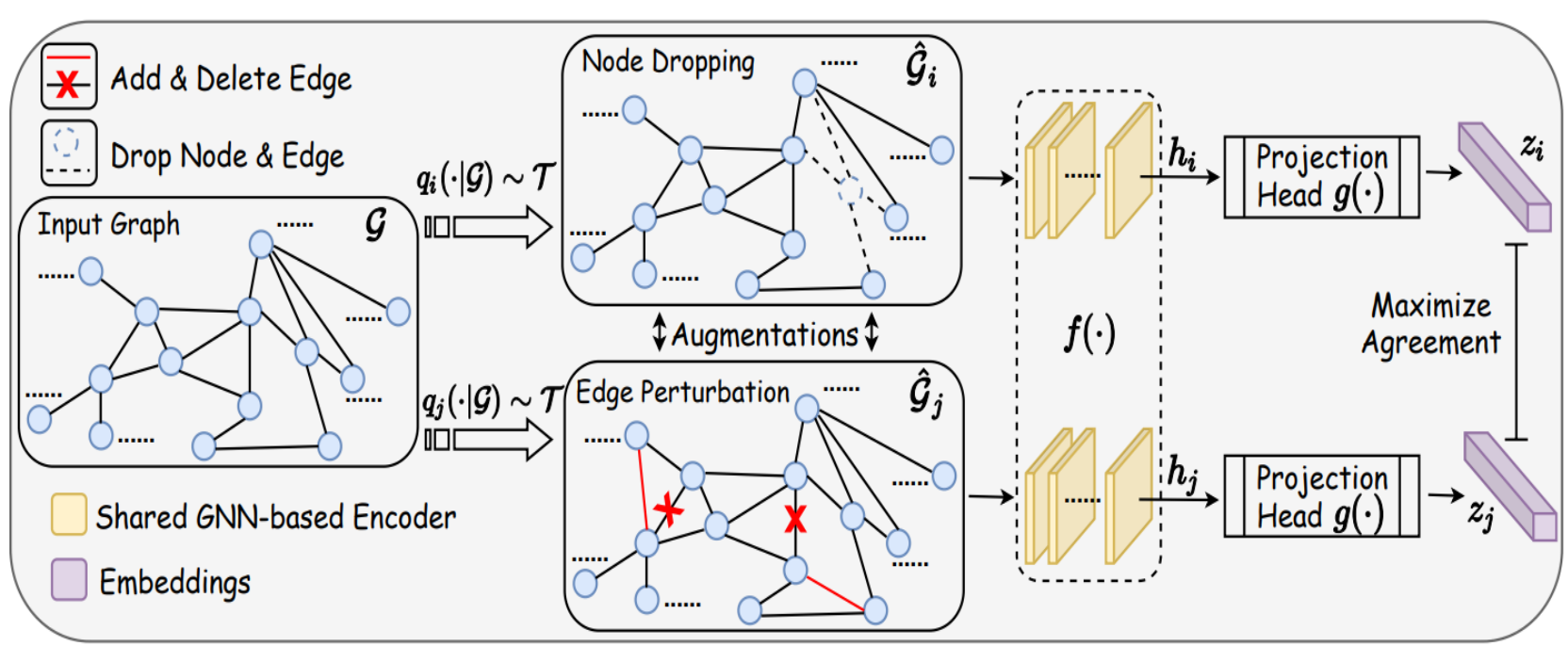
Finally, Use **Tiny BERT** for the next step embedding.

Tiny BERT finetune architecture



BERT + GCL Pretraining

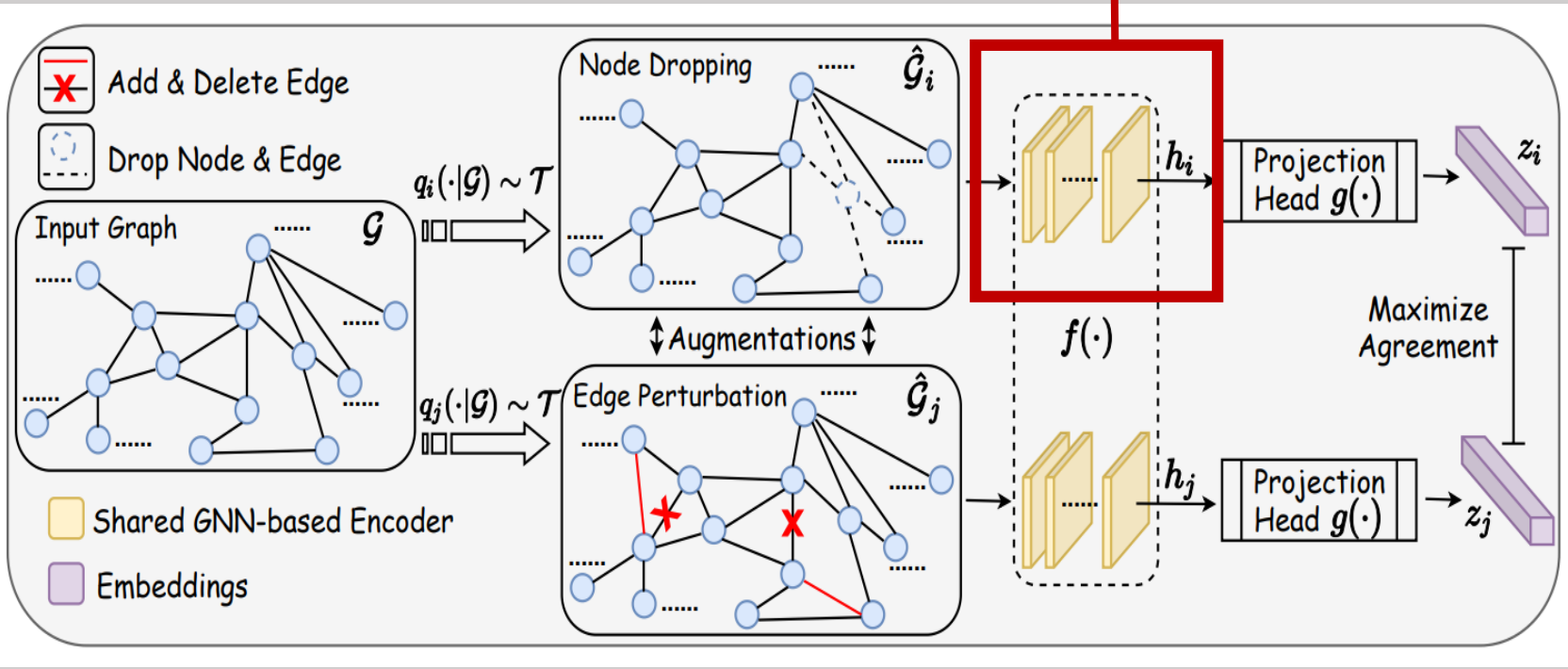
Node: Authors/Title/Date → **Tiny BERT** → **GraphCL** → Contrastive Loss



GraphCL architecture

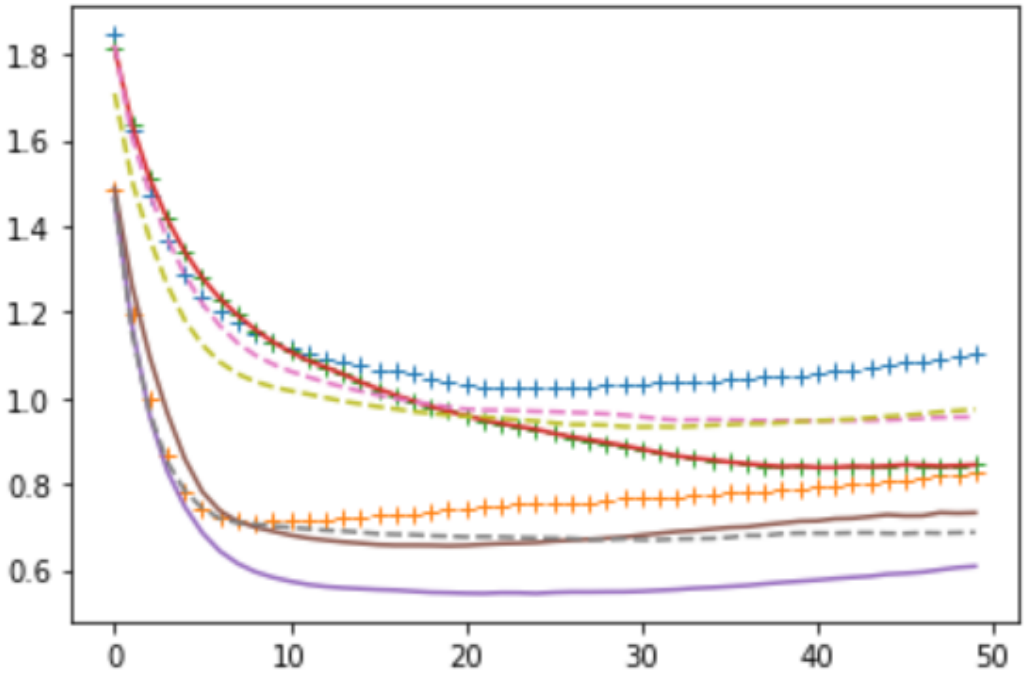
BERT + GCL Finetune

Node: Authors/Title/Date → **Tiny BERT** → **GraphConv** → **MLP** → BCE Loss



GraphCL architecture

Choose Pretraining mode



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

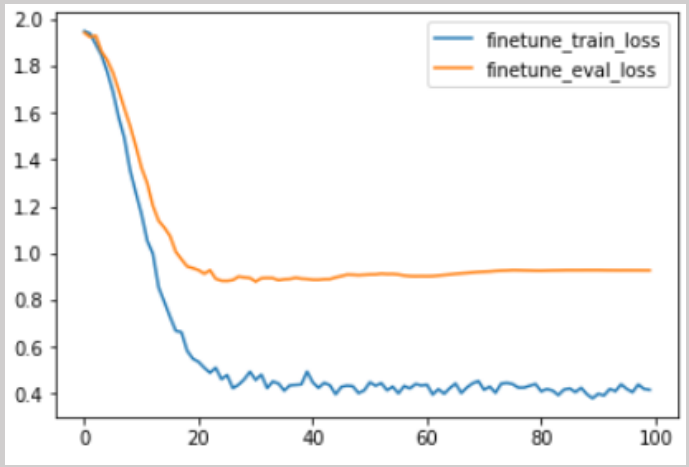
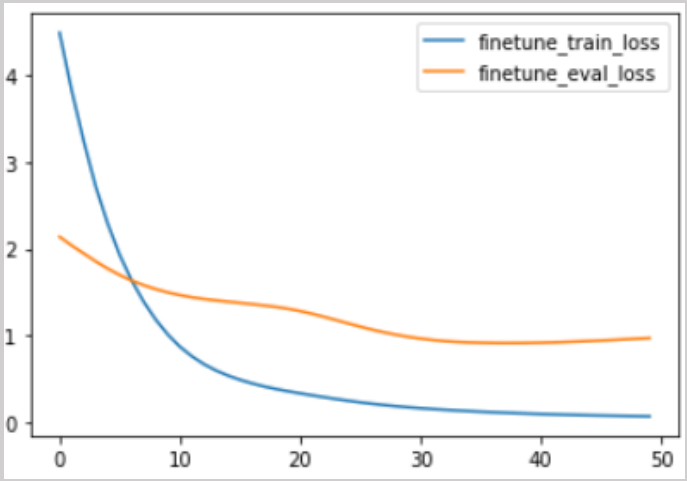
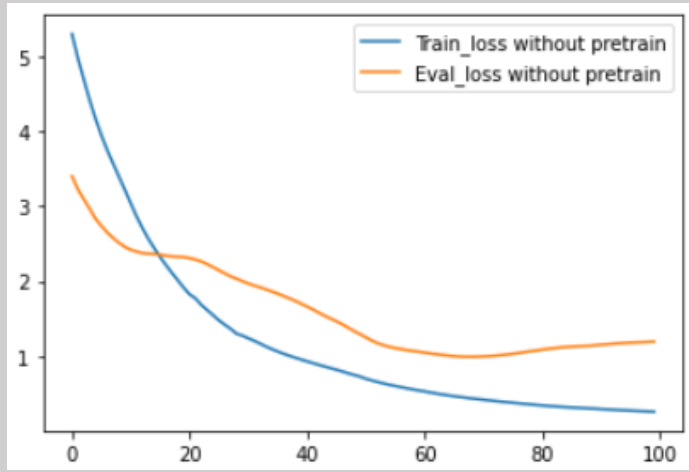
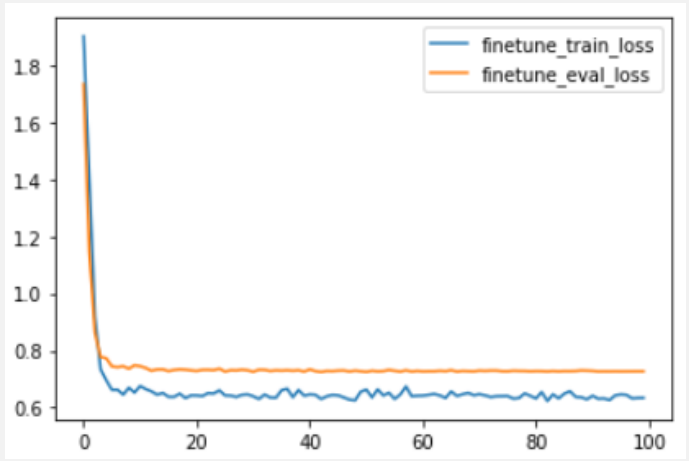
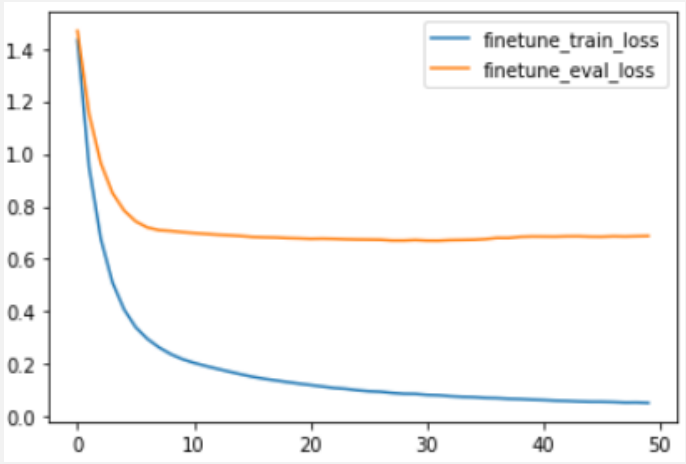
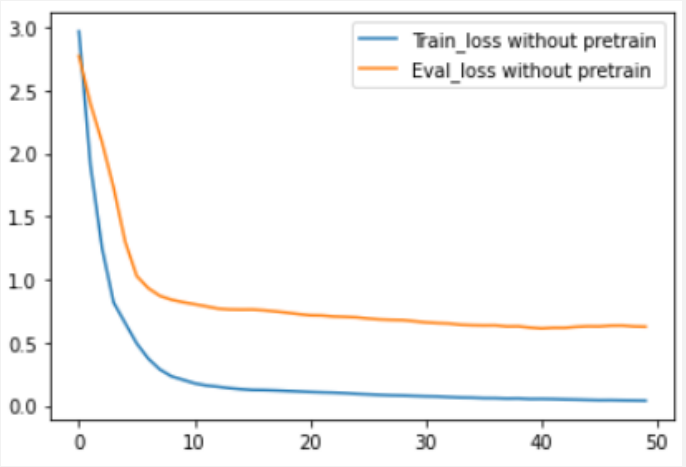
紅色代表: 使用這個augmentation graph的GCN做finetune

Difference

GraphCL	(GCN + BatchNorm) *3 + (Linear + relu) *1 + Linear	(GCN + BatchNorm) *2 + (Linear + relu) *1 + Linear **My Model**

Others: GCL not freeze (GCN + BatchNorm) in original paper

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

	BERT	BERT + GCL (Pretraining : RW + NodeDrop)	BERT + GCL (without Pretraining)
Label : 0.01%			
	Accuracy:0.690	Accuracy:0.626	Accuracy:0.628
Label : 0.1%			
	Accuracy:0.705	Accuracy:0.841	Accuracy:0.811

Discussion

GNN-based model is suitable for similar journals prediction

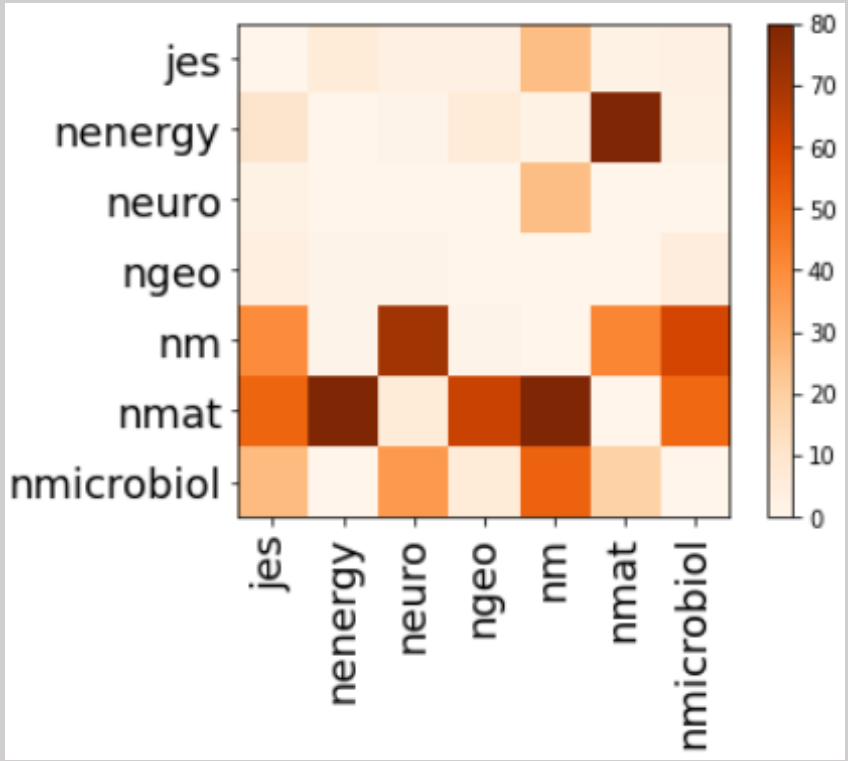
Label Rate	Tiny BERT (date/title/author)	BERT (abstract)	GraphCL (one hot)	Bert + GCL	Bert + GCL (without Pretrain)
1%	0.690	0.663	0.300	0.626	0.628
10%	0.705	0.788	0.403	0.841	0.811

Error Prediction

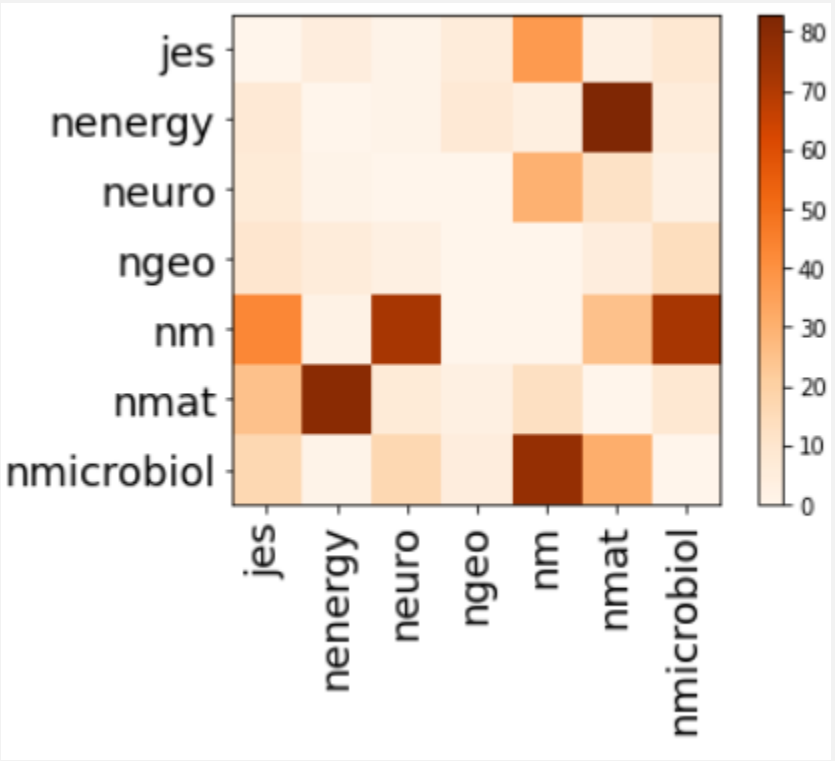
Without pretrain

Amount of Data

jes: 900
nenergy: 570
neuro: 747
ngeo: 823
nm: 652
nmat: 777
nmicrobiol: 843



pretrain



Prediction

Actual

Prediction error

Prediction: nature energy / Actual: nature material

1. benduhn, j. et al. intrinsic non - radiative voltage losses in fullerene - based organic solar cells. nat. energy 2, 17053 (2017).

Prediction: nature medicine / Actual: nature neuroscience

2. chen, j. a., penagarikano, o., belgard, t. g., swarup, v. & geschwind, d. h. the emerging picture of autism spectrum disorder : genetics and pathology. (2015).

Prediction: nature medicine / Actual: nature microbiology

3. li, y. et al. exogenous stimuli maintain intraepithelial lymphocytes via aryl hydrocarbon receptor activation. (2011).

Prediction: nature material / Actual: nature energy

4. 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).

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.