KDD CUP 2021 MAG240M-LSC TEAM PASSAGES WINNER SOLUTION

A PREPRINT

Kaiyuan Li[*], **Xiang Long**[*], **Zhichao Feng**[*], **Mingdao Wang**[*], **Xiaofan Liu**[*], **Pengfei Wang** [*]
Beijing University of Posts and Telecommunications
Beijing, China
{tsotfsk, xianglong, guiji_ql, wmingdao, liuxiaofan, wangpengfei}@bupt.edu.cn

Quan Lin [†]
Huazhong University of Science and Technology
Wuhan, China
iviolet4ever@gmail.com

Kun Zhao [‡], **Bole Ai** [‡]
Nanjing University
Nanjing, China
{jackon.zhao, baoleai01}@gmail.com

June 16, 2021

ABSTRACT

OGB LSC@KDD Cup 2021 proposes a Multi-class classification task on MAG240M-LSC dataset, in which the participants are asked to predict the subject areas of papers situated in a heterogeneous academic graph. The metric is the classification accuracy. This paper describes our winner solution on MAG240M-LSC dataset, which can be largely summarized in two main stages: a pretraining stage is designed to explore heterogeneous academic networks for better node embeddings; and a transfer learning stage is used to alleviate differences of label distributions and node representations between training and test set. Our solution achieves an overall score of 0.738.

Keywords KDD Cup · Multi-class Classification · Heterogeneous Academic Graph · Transfer Learning · Pretraining

1 Introduction

KDD Cup 2021 OGB Large-Scale Challenge proposes a large-scale graph ML competition to encourage the development of state-of-the-art graph ML models for massive modern datasets. In this track, the task on MAG240M-LSC dataset is to predict the subject areas of papers. Specifically, MAG240M-LSC is a heterogeneous academic graph extracted from the Microsoft Academic Graph, it contains 121M English academic papers, 122M authors, and 26K academic institutions. These relations among these nodes are offered, including paper-cites-paper, author-writes-paper, and author-affiliated-with-institution. Given this datatset, we aim to automatically annotate paper topics, i.e., predicting the primary subject area of each paper.

Our solution for the node classification task contains two important stages: 1) We proposed one graph model to pretrain node embeddings under different tune tricks, and ensemble them to capture complicated interactions among these heterogeneous nodes. 2) We found label distributions and paper representations in adjacent years are similar, while in nonadjacent years, they show strong variations. Consider the large time gap between training set and test set, we use the model learned on validation set for test, and a transfer learning approach is applied to tune this model based on the one learned from training set. Our solution achieves an overall score of 0.738.

[*]Beijing University of Posts and Telecommunications
[†]Huazhong University of Science and Technology
[‡]Nanjing University

2 Method

In this section, we introduce two major stages to enhance the prediction performance, including the pretraining stage for better representations, and a transfer learning stage to bridge data gaps between training set and test set. Specifically,

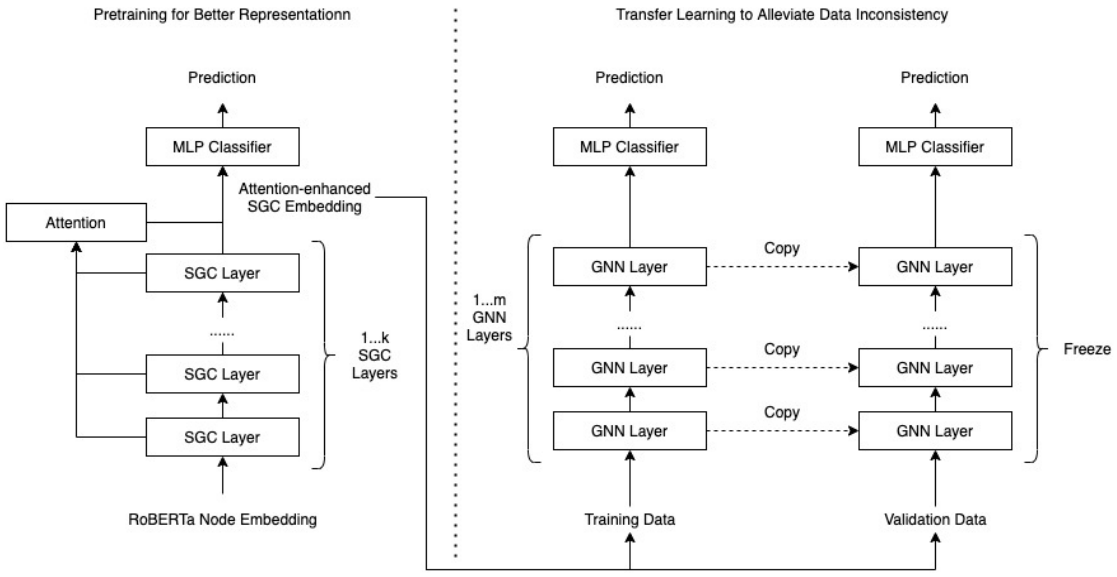

Figure 1: Total pipeline of our solution.

in the pretraining stage, we use SGC to obtain enhanced paper embeddings. Given the paper embeddings learned by SCG and RoBERTa, we then feed them into R-GAT respectively for classification. In the transfer learning stage, for each type of paper representations, we learn the R-SGAT according the training set, and utilize the model learned from training set to tune the one obtained on validation set to alleviate data inconsistency between training set and test set. A late fusion based on voting technique is used to aggregate them as the our final solution. The structure of our solution is shown in Fig. 1.

2.1 Pretraining for Better Representation

The pre-training on the graph is an meaningful metric for this task, because paper embeddings are made by RoBERTa [Liu et al., 2019] language model, which only considers paper textual information, while paper citations, authors, and institutions are all ignored. In addition, only 1% of papers are labeled. To fully utilize heterogeneous information in the academic graph, we first apply pretrain metric to obtain better node representations.

However, when applying pre-training on large scale graphs, it needs large requirement of the storage and more computing time. We first consider to use some shallow pre-training schemes such as TransE[Bordes et al., 2013], metapath2vec[Dong et al., 2017], deepwalk[Perozzi et al., 2014], and test their performance on a small sampled dataset. However, when running these models on MAG240M-LSC, we found these algorithms are difficult to work under limited hardware conditions. Therefore, how to generate a simple pre-training model to fully aggregate heterogeneous information becomes the first key point we need to address.

Recently, SGC [Wu et al., 2019] has been proved to have similar performance as GCN [Kipf and Welling, 2016], and it can well address both sparse matrix and large matrix multiplication even in limited hardware resources, which is quite fit for this task. Therefore, we try to use SGC to pre-train our model, and make some improvement for learning. Specifically, we first use SCG to calculate the k-order laplacian matrix over papers according their citation relations, and obtain its corresponding paper representations. An attention metric is further used to aggregate them over different laplacian matrices as the final enhanced paper representations.

After this, for each paper, we then have its initial representation obtained by RoBERTa, and enhanced representation learned by the retraining approach. Both of them are further feed into R-GAT for classification.

2.2 Transfer Learning to Alleviate Data Inconsistency

In this issue, we found label distributions and node representations are similar in adjacent years, while in nonadjacent years they show strong variations. We illustrate both effects by following two figures. Figure 2 shows label distribution similarities from 2010 to 2019. Figure 3 shows the average node semantic embedding shifting by year.

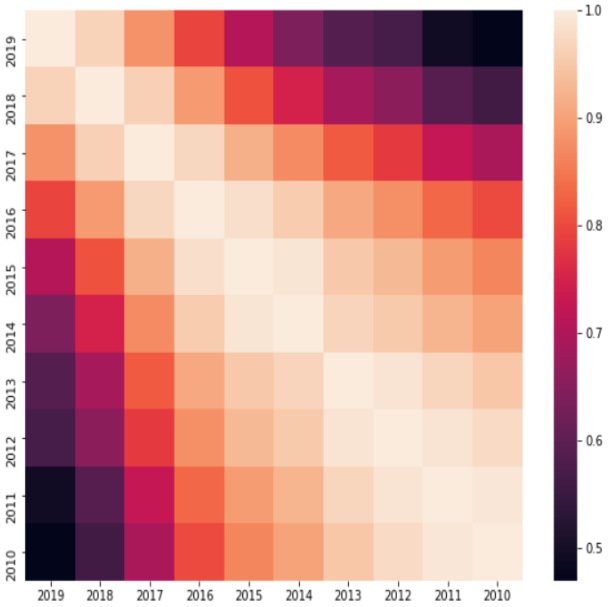

Figure 2: Heatmap of Label Similarities 2010-2019

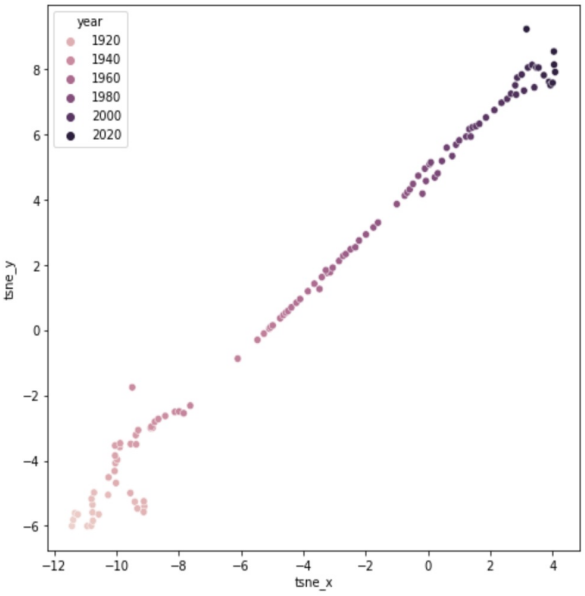

Figure 3: Node Embedding Shift by Year

Due to the data inconsistency between training set and test set, directly applying the model learned from training set for test may impact the prediction performance. As publish time of papers on validation set and test set are adjacent, ideally, learning our model on the validation set for test seems a promising approach. But in our experiments, due to its limited data, directly learning our model on validation set for test could not work well. How to well utilize all data information to learn a better model on validation set for test thus becomes another key point we need to address.

To address this point, we first use the training set to train R-GAT for classification based on the paper representations. We then fix the parameters of the GNN layers of R-GAT and re-train its MLP layers on validation set to fit the data. We repeat the above process on both initial and enhanced paper representations to obtain two fine-tuned models for test, and utilize late fusion metric to aggregate their results as our final solution.

3 Experiments

In this section, we further conduct a series of detailed experiments to analyze the effectiveness of our approach.

3.1 Parameter Setting

For each node of R-GAT, we sample at most 25 neighbors in the first layer, and 15 neighbors in the seconder layer on training set, and sample 200 neighbors for two layers on validation set. At inference time, we sample at most 200 neighbors for each layer. For Both SGC and R-GAT, the hidden dimensionality is set to 2048, and propagation layers for SGC is set to 3.

3.2 Experiment Results

In this section, we give the performance of our model when applying different model ensemble and tune strategies in pretraining stage and transfer learning stage.

3.3 Transfer Learning Strategies

We give several variations of our tune strategies to analyze their performance:

- **None-tune strategy**: we use training set to train our R-GAT, and test its performance on validation set.
- **Rough-tune strategy**: For GNN layers of R-GAT, we copy parameters that learned under none-tune strategy and fix them. For the MLP layer, we randomly initialize them, and tune them according 5-fold cross validation on validation set.
- **Fine-tune strategy**: we copy all parameters that learned under non-tune strategy, and tune MLP layer according 5-fold cross validation on validation set.

We consider model performance when utilizing different paper representations: initial representations obtained by RoBERTa (denoted as RoBERTa+R-GAT), and the enhanced representations learned by SGC (denoted as SGC+R-GAT). The result is shown in Table 1.

Table 1: Performance comparison of models under different tune policies.

Model	None-tune	Rough-tune	Fine-tune
RoBERTa+ R-GAT	0.7064	0.7294	0.7355
SGC+R-GAT	0.7081	0.7293	0.7359

3.4 Model Ensemble Strategies

In this section, we further analyzed performance of our model when applying different ensemble strategies. We consider the following ensemble strategies:

- **Voting**: The architecture of a Voting Classifier is made up of a number "n" of ML models, whose predictions are valued in two different ways: hard and soft. In hard mode, the winning prediction is the one with "the most votes". On the other hand, the Voting Classifier in soft mode considers the probabilities thrown by each ML model, these probabilities will be weighted and averaged, consequently, the winning class will be the one with the highest weighted and averaged probability.
- **Stacking**: Better known as Stacking Generalization, it is a method introduced by David H. Wolpert in 1992 **??** where the key is to reduce the generalization error of different generalizers (i.e. ML models). The general idea of the Stacking Generalization method is the generation of a Meta-Model. Such a Meta-Model is made up of the predictions of a set of ML base models (i.e. weak learners) through the k-fold cross-validation technique. Finally, the Meta-Model is trained with an additional ML model (which is commonly known as the "final estimator" or "final learner").

The performance result is shown in Table 2.

Table 2: Performance of our model under different Model Ensemble metrics

Ensemble Metric	Validation set	Test set
Voting-Hard	0.7413	-
Stacking	0.7427	-
Voting-Soft	0.7440	0.7381

3.5 Conclusion

In this paper, we present our solution for the KDD Cup 2021 OGB-LSC MAG240M-LSC competition. We propose a set of feasible methods in the current real-world large-scale graph data scenario, including SGC-like pre-training methods to enhance node features, graph representation learning in a message-passing framework, MLP classification of node representations aggregating nearest neighbor information, and migration learning-based methods to mitigate the performance degradation caused by different training-validation-test set distributions. By discussing the results under different settings, we demonstrate the effectiveness of the methods and achieve the top three positions in the competition.

References

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 2787–2795, Red Hook, NY, USA, 2013. Curran Associates Inc.

Yuxiao Dong, Nitesh V. Chawla, and Ananthram Swami. Metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 135–144, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450348874. doi:10.1145/3097983.3098036. URL https://doi.org/10.1145/3097983.3098036.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, page 701–710, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450329569. doi:10.1145/2623330.2623732. URL https://doi.org/10.1145/2623330.2623732.

Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Weinberger. Simplifying graph convolutional networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6861–6871. PMLR, 09–15 Jun 2019. URL http://proceedings.mlr.press/v97/wu19e.html.

Thomas Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. 09 2016.