Open Graph Benchmark - Large Scale Challenge @ KDD Cup 2021

Master Project

Group D

July 12, 2021

Kiel University
Department of Computer Science

Advised by: Christian Beth
Andreas Lohrer

Contents

Chapter 1

Big Picture

The aim of this master project was to take part in the KDD cup challenge which is hosted by OGB. The idea was to build a graph neural network model for the MAG240M-LSC dataset. It was a node classification task where the model was supposed to predict the field of study for any given paper.

1.1 Motivation and Goals

Graph data is everywhere. As an example Facebook, Twitter use graph to predict the friend suggestions and advertisement. Graph data is so complex that it's created a lot of challenges for existing machine learning algorithms. The reason is that conventional Machine Learning and Deep Learning tools are specialized in simple data types. Like images with the same structure and size, which we can think of as fixed-size grid graphs. Text and speech are sequences, so we can think of them as line graphs. But there are more complex graphs, without a fixed form, with a variable size of unordered nodes, where nodes can have different amounts of neighbors. All the nodes occupy an arbitrary position in space. It also doesn't help that existing machine learning algorithms make an assumption that instances are independent of each other. This is not true for graph data, because each node is related to others by links of various types. The main goals are

- Learning about graph neural network
- KDD Cup 2021
- Apply GNN on heterogeneous graph

1.2 Steps of Project

Tasks	Responsible Persons
Literature Research	Faiz Ahmed, Abdullah Al Amin, Mithun Das, Rishabh Lakra, Md Mashiur Rahman,Shaokat Hossain, Ankit Malhotra, Md Abu Noman Majumdar
Learn/Implement Homo/Heterogeneous graph	Faiz Ahmed, Abdullah Al Amin, Mithun Das, Md Mashiur Rahman,Shaokat Hossain, Ankit Malhotra, Rishabh Lakra, Md Abu Noman Majumdar
MAG240M-LSC (200 GB) Sampling and Visualizing	Faiz Ahmed, Abdullah Al Amin,Shaokat Hossain, Mithun Das, Md Mashiur Rahma,Md Abu Noman Majumdar
OGBN-MAG (1GB)	Faiz Ahmed, Abdullah Al Amin, Mithun Das, Md Mashiur Rahman, Shaokat Hossain, Ankit Malhotra, Rishabh Lakra,Md Abu Noman Majumdar
ACM Dataset	Abdullah Al Amin, Md Mashiur Rahman,Mithun Das
Presentation and Report	Faiz Ahmed , Md Mashiur Rahman, Mithun Das,Md Abu Noman Majumdar , Rishabh Lakra, Ankit Malhotra

1.3 Results and Findings

Although We aimed to participate in the KDD cup we were not able to work the the 200 GB graph(Microsoft Academic Graph) because of hardware limitations and also the limitation of the Deep Graph Library. Then we decided to work with the 2GB variant of the MAG. There we were able to make one model but we couldn't proceed further because of the same hardware and library issues. It seems the DGL is not yet ready to be applied to big graphs and there are open discussions about these feature requests.Finally we decided to work with the ACM dataset since it is structurally very similar to the MAG dataset. The task was to predict the venues of each paper(link prediction). The results from the experiment is shown below.

Model	Loss(Cross_entropy)	Train_Acc	Test_Acc	Total_Epoch
HeteroRGCN	0.41787204146385193	0.9241666793823242	0.7811447978019714	20
HAN	0.3301704227924347	0.8716666666666667	0.8400673400673401	20
HeteroRGCN	0.11036919802427292	0.9950000047683716	0.9343434572219849	40
HAN	0.19656327366828918	0.9333333333333333	0.867003367003367	40
HeteroRGCN	0.03761648014187813	1.0	0.9680134654045105	60
HAN	0.1779087334871292	0.9366666666666666	0.8602693602693603	60
HeteroRGCN	0.018344396725296974	1.0	0.9764309525489807	80
HAN	0.13779298961162567	0.9558333333333333	0.8602693602693603	80
HeteroRGCN	0.014451955445110798	1.0	0.9797979593276978	100
HAN	0.12297426909208298	0.9608333333333333	0.8552188552188552	100
HeteroRGCN	0.01277798693627119	1.0	0.9831649661064148	120
HAN	0.10522094368934631	0.9641666666666666	0.8585858585858586	120
HeteroRGCN	0.009530107490718365	1.0	0.9898989796638489	140
HAN	0.12542876601219177	0.9475	0.8569023569023569	140

Chapter 2

Personal Achievement and Contribution

2.1 Abdullah Al Amin

- Learnt about the basics of homogeneous graphs and heterogeneous graphs
- Studied the research paper about an attention based GNN which is used for Heterogeneous structural learning. The idea of this research was to apply an attention based model which will learn about the underlying structure of a given heterogeneous graph, and by doing so It can work efficiently on the structure which is addressed by metapaths in other algorithms. But the metapaths are highly manual effort intensive.
- Learnt some of the useful API's from the Deep graph library(DGL) which are used for building graph neural networks as well as creating and manipulating heterogeneous graphs. Although DGL is a very nice library with high level of abstractions(it is built upon the pytorch geometric) it is currently under development and not optimized for big graphs.
- Worked on a very naive sampling technique for our initial 200 gb dataset. The idea was to just select a portion of the nodes and their interconnecting edges in a sequential manner. There is an API for this as well in the DGL but it didn't work out for the big data because of the i outdegree edges of the selected edges recursively selects the nodes connected and it goes on. So in the end we end up with a huge data. There are researches about this topic but they are very new and not tested in the field. Also The sampling of a heterogeneous graph is non trivial task. It involves preserving the in-lying distribution as well as the heterogeneous semantic information of the whole graph.
- Wrote a RGCN(Relational Graph Convolutional Network) model for the 2gb variant of the MAG(Microsoft Academic Graph). The graph only has features for one kind of nodes(paper). So in order to work with the missing data I first filled the other features with zero values. And for another variant I tried to use the featureless embedding(the paper feature was also disregarded) and the embedding was initialized with Xavier uniform initializer. The first setting achieved 28 percent accuracy on the data but the second setting could not be run on the paper space.
- Wrote a HAN(Heterogeneous graph Attention network) for the ACM graph. This model achieved 86 percent accuracy on the ACM graph dataset.

▻ Combined the RGCN model written for ACM data with the HAN model. Wrote an test setting script with a comprehensive logger mechanism. This script can be used to run and recreate the experiment on the ACM graph that we did for the benchmark analysis.

▻ Wrote an analysis script that works on the experiment of the HAN and RGCN model data and creates some visualization for benchmark comparison.

2.2 Ankit Malhotra

▻ Understanding the basic concepts like the meaning of convolution in a graph, concept of attention in Graph Neural Networks.

▻ Understanding the APIs in DGL like representation of nodes and edges, manipulation of data, etc.

▻ Implemented changes in the OGBN-MAG dataset i.e for the proper working of the given HAN model the reverse edges (from destination to source) were to be added in the given heterogeneous graph but the inbuilt APIs only supported this functionality for homogenous one.

▻ Implemented changes in the baseline HAN model to make it work on the OGBN-MAG dataset but due to the size of the dataset and lack of computing resources the full batch training was not possible, an out-of-memory error was thrown.

▻ Worked on implementing stochastic training for the OGBN-MAG dataset, stochastic training on graphs is done by defining a neighborhood sampler, adapting your model for minibatch training, and then modifying the training loop. We then use a node data loader that iterates over a set of nodes in mini-batches. The problem that was encountered here was the HAN model defines neighborhoods based on meta path. However, this step makes the graph a lot denser. Moreover, the meta-path reachable graph API runs on the CPU.

2.3 Faiz Ahmed

Before this Master project I even didn't know about Heterogeneous Graph. But, I did know that graph data is everywhere and working with graph data will be interesting, that's why I choose this Master project. If didn't consider the infrastructure limitation, I can easily say that now I can handle the large scale heterogeneous graph data, which I think a big achievement for me. To achieve this goal the following things I did so far:

- Literature Research: Homo/Heterogeneous graph, Image Convolution and Graph Convolution, GNN, GCN, Message passing Framework, GAT. Git link for the presentation of initial literature study: `https://git.informatik.uni-kiel.de/ag-isdm/modules/ml-master-projects/kdd-cup2021/dgl-test/-/blob/master/presentation.pdf`
- Implementation of Heterogeneous graph. `https://git.informatik.uni-kiel.de/ag-isdm/modules/ml-master-projects/kdd-cup2021/toy-codes-faiz/-/blob/master/zacharys-karate-club-with-dgl.ipynb`
- Used some API's of DGL
- Played with **MAG240M-LSC (200 GB)**. Tried to sub sample and process this dataset. But failed for infrastructure limitation. Some code links:
  - `https://git.informatik.uni-kiel.de/ag-isdm/modules/ml-master-projects/kdd-cup2021/toy-codes-faiz/-/blob/master/MAG240MDataset.ipynb`
  - `https://git.informatik.uni-kiel.de/ag-isdm/modules/ml-master-projects/kdd-cup2021/toy-codes-faiz/-/blob/master/graph_subsampling.ipynb`
  - `https://git.informatik.uni-kiel.de/ag-isdm/modules/ml-master-projects/kdd-cup2021/toy-codes-faiz/-/blob/master/preprocess.py`
- GAT implementation : `https://git.informatik.uni-kiel.de/ag-isdm/modules/ml-master-projects/kdd-cup2021/toy-codes-faiz/-/tree/master/gat`
- Worked also with **OGBN-MAG (1GB)** , but again RAM limitation
- Helped in Presentation preparation and final report.

2.4 Md Abu Noman Majumdar

- Learning about classification , prediction and clustering of homogeneous and heterogeneous graph data through graph neural network.
- Studied research papers about Graph convolution Network, Graph Attention Network , Graph Sage Network and Recurrent Graph Convolution Network.
- Learnt about Pytorch and Deep graph library(DGL) and some useful API's of DGL to implement our model.

- ▻ For MAG240M-LSC (200 GB) have learnt Neighborhood sampling for minibatch graph generation though we couldn't implement this as it is not trivial task and also learnt the basic concept of distributed training in DGL and learend two partitioning algorithm 'Metis' and 'Random'.
- ▻ Worked and discussed in data pre-processing and implementation of model with my group member.
- ▻ Worked on making final presentation .

2.5 Md. Mashiur Rahman

- ▻ Learned about the concept of Graph and it's types like homogeneous graphs and heterogeneous graphs.
- ▻ Studied research papers based on Graph Convolution, GNN and attention based GNN to gather the underlying concept and idea about how GNN works on heterogeneous graph and learned some useful API's from the Deep graph library(DGL).
- ▻ Regarding MAG240M-LSC dataset, although heterogeneous graphs sampling is a big challenging task, i tried to handle sampling of heterogeneous graphs with help of my colleagues sampling method in the PaperSpace machine. But, it didn't work out the way we wanted due to resource limitation in the PaperSpace and some limitations in the DGL library.
- ▻ Regarding OGBN-MAG Dataset, combined both RGCN (Prefilling missing feature with zero value) and HeteroRGCN (Featureless Embedding) models and Train both models for Node Classification task with CSV file logger.
- ▻ Regarding ACM Dataset, wrote a RGCN (based on Featureless Embedding) model for Node Classification task with CSV file logger. This model achieved 97 percent accuracy.
- ▻ As responsible person for PaperSpace machine, i worked with all 3 DataSets. I did observe, experiment, execute, train relevant models and run relevant scripts based on all 3 datasets and gather the findings which datasets and which model variant didn't work properly and the reasoning behind it.
- ▻ Participated all three (research paper group-wise presentation, half-time project presentation and final project presentation) presentations as presenter.

2.6 Mithun Das

At the beginning of the master project, I had knowledge about deep learning and machine learning, but I had zero idea about Graph neural network.It took me a while to completely part of the team as I had to study a lot of topics, a lot of frameworks that related to graph neural network, Had to understand the implementation technique of the API that provided by DGL. I support my team mostly on researching most efficient technique for implementation and implementing those while needed.In the below section, I will note down my whole journey of the master project...

- Gain an understanding of homogeneous graphs and heterogeneous graphs by studying the related research paper given by our supervisor and from internet as well.
- Understand the concept of GNN, GAT, Image convolution. Played around with the all the APIs in DGL that we needed to use by implementing in small examples for practicing the API response.
- Researched about all the possible idea of sampling a Large data, we had **MAG240M-LSC (200 GB)**.
- At one moment when we found that sampling this size of data need higher computational machine, I started searching for a similar dataset which has the exactly features as **MAG240M-LSC (200 GB)**.
- I proposed these 3 dataset which I got finally in kaggle and 2 in OGB dataset itself to my team.It has almost same features but in every dataset we had different task to do. But finally we took **OGBN-MAG (1GB)** which has same feature as the **MAG240M-LSC (200 GB)** dataset.My proposed dataset are below:
  - `https://www.kaggle.com/dataup1/ogbn-arxiv`
  - `https://www.kaggle.com/dataup1/ogbn-mag`
  - `https://www.kaggle.com/dataup1/ogbg-moltoxcast`
- After selecting the lower size dataset (**OGBN-MAG (1GB)**), my main task was to search for some large dataset sampling technique. There were few of them from which I selected this one to apply as this one is highly rated on GitHUb and had great response by the community.
  - `https://little-ball-of-fur.readthedocs.io/en/latest/modules/node_sampling.html`
- At the beginning I got total 4 frameworks that might help use doing the sampling of the whole data set.After fitting all of those in the dataset, I tried Little-ball-of-fur framework to do sampling of the **OGBN-MAG (1GB)** dataset here..
  - `https://git.informatik.uni-kiel.de/ag-isdm/modules/ml-master-projects/kdd-cup2021/dgl-practice/-/tree/dgl-practice-md`

- Worked on implementing for the **OGBN-MAG (1GB)** dataset to node classification on mag-checkpoints.
  - https://git.informatik.uni-kiel.de/ag-isdm/modules/ml-master-projects/kdd-cup2021/dgl-practice/-/blob/dgl-practice-md/dataset/.ipynb_checkpoints/heterograph%20node%20classification%20mag-checkpoint.ipynb
- Prepared the Final presentation with the help of whole team.

2.7 Rishabh Lakra

- Covered a lot of territory in the course of working on this project, and gained useful insight in the process of implementing the different models. Learning about graph data and graph neural networks was the first step in the project, which was achieved by reading the recommended research papers at the beginning of the project. It was a fairly new concept, so it was clear going forward will be challenging but there will be a lot to learn about a promising advancement in the field of data mining. The research papers gave a good start to dive into the world of GNNs and other advanced models like GCNs, R-GCNs, GATs, and especially HANs, where understanding the attention mechanism was the first step.
- Also, throughout the whole project, the Deep Graph Library was constantly a source for learning more about how this kind of data can be handled and exploited by a graph neural network and spent considerable time on reading the DGL documentation and learning APIs that are relevant for the project. Although later on in the project it turned out to be insufficient for the scale and complexity of the heterogeneous dataset we were presented with.
- Based on the above, implemented a graph neural network on small-scale movie data to understand better the graph data and to see if this can be translated further to our main model.
- From here, the work towards realizing the first goal started, which was the KDD cup. The aim was to develop a HAN model and applying it to the hetero-graphs presented by the MAG dataset. The dataset itself was a challenge as it caused scalability and computational issues, as we did not have machines capable enough to handle it. And by now it was clear that heterogeneous graph data has a very complex structure and is semantically rich, but these characteristics were seemingly untameable at the beginning.
- Tried working on a subset of the data when the above did not work. And re-ran the model, debugging, finding issues, and solving them was a constant task. Also, there were a lot of iterative tasks in trying different solutions to tackle the issues we faced.

- A lot of effort was required to research, build and fine-tune a HAN model. Even for a student research project like the one we built, it took an immense amount of labor to pre-process, filter, and organize the data before any algorithms could run.
- Lastly, helped in preparing and delivering the final presentation.

2.8 Shaokat Hossain

- Studied some research papers on homogeneous and heterogeneous graphs. Specially a research paper about graph attention network named Heterogeneous Graph Attention Network, where a novel heterogeneous graph neural network based on hierarchical attention, has been proposed. The model follows a hierarchical attention structure: node-level attention(learn the importance between a node and its meta-path based neighbors)→semantic-level attention(learn the importance of different meta-paths). The HAN has good interpretability for learning node embedding which is a big advantage of heterogeneous graph analysis. The proposed model leverages node-level attention and semantic-level attention to learn the importance of nodes and meta-paths, respectively. Based on the importance values we can check the higher and lower contributed nodes or metha paths. By doing all of these it provides superior performance.
- Studying Deep Graph Library(DGL) and pytorch geometric. I came across DGL while doing the project and got the opportunity to learn some API's. We found DGL is not very feasible for doing big graphs(still under development), though it has very nice API's for creating and manipulating Heterogeneous graphs.
- Doing research on distributed training to overcome the infeasible hardware infrastructure on Paperspace for our massive amount of graph dataset. Unfortunately we didn't have enough knowledge to run the model on a distributed machine and the time frame was also short.
- For selecting the best model I wrote a GCN and GAT model and applied coradataset on it, in GAT at first got very less accuracy then optimised the model by adding GATConv and got good accuracy.
- I did data preprocessing for our project. Throughout our project we learnet and got familiar with DGL but mysteriously with DGL loader we didn't get any dataset but graph structure for OGBN-MAG dataset . Then I used Library-Agnostic Loader to get the dataset but it has to add in DGL supported graph and also missing mask in dataset.
- Attending this project, I have learnt handling large scale graph dataset building models for heterogeneous graphs. Apart from this it helped me to improve my skills to work in a team.