Multi-Dataset and Contrastive Learning on GIANT-XRT

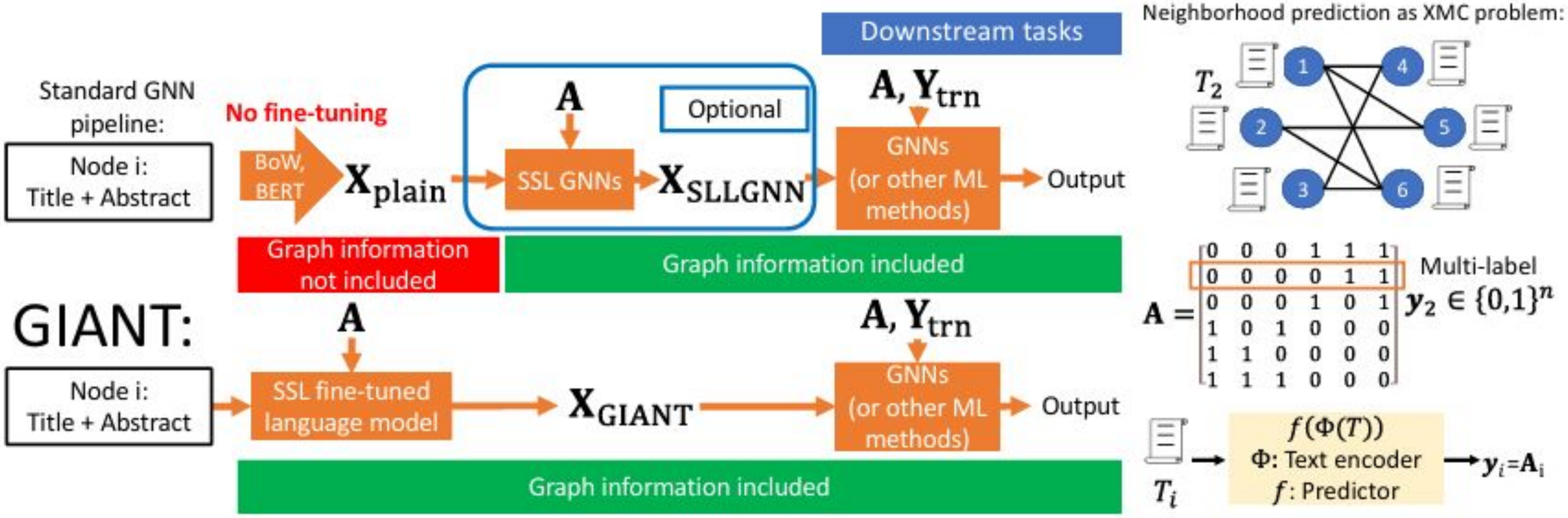
Wenhao Xu (wx2056)

Text-Attribute GNN

Previous research has focused on the relationship between numerical node features and graph structure, enhancing GNN performance. However, existing methods for extracting these features remain **graph-agnostic**, hindering the utilization of graph-topology correlations.

GIANT-XRT, a self-supervised learning framework leveraging XMC and XR-Transformers for improved performance on large datasets.

GIANT-XRT


Standard GNN pipeline:
Node i: Title + Abstract
No fine-tuning
BoW, BERT
$\mathbf{X}_{\text{plain}}$
A
SSL GNNs
Optional
$\mathbf{X}_{\text{SLLGNN}}$
Downstream tasks
$\mathbf{A}, \mathbf{Y}_{\text{trn}}$
GNNs (or other ML methods)
Output
Graph information not included
Graph information included
GIANT:
Node i: Title + Abstract
A
SSL fine-tuned language model
$\mathbf{X}_{\text{GIANT}}$
$\mathbf{A}, \mathbf{Y}_{\text{trn}}$
GNNs (or other ML methods)
Output
Graph information included
Neighborhood prediction as XMC problem:
T_2
1
2
3
4
5
6
Multi-label
$\boldsymbol{y}_2 \in \{0,1\}^n$
T_i
$f(\Phi(T))$
Φ: Text encoder
f: Predictor
$\boldsymbol{y}_i$=$\mathbf{A}_i$

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Limitation

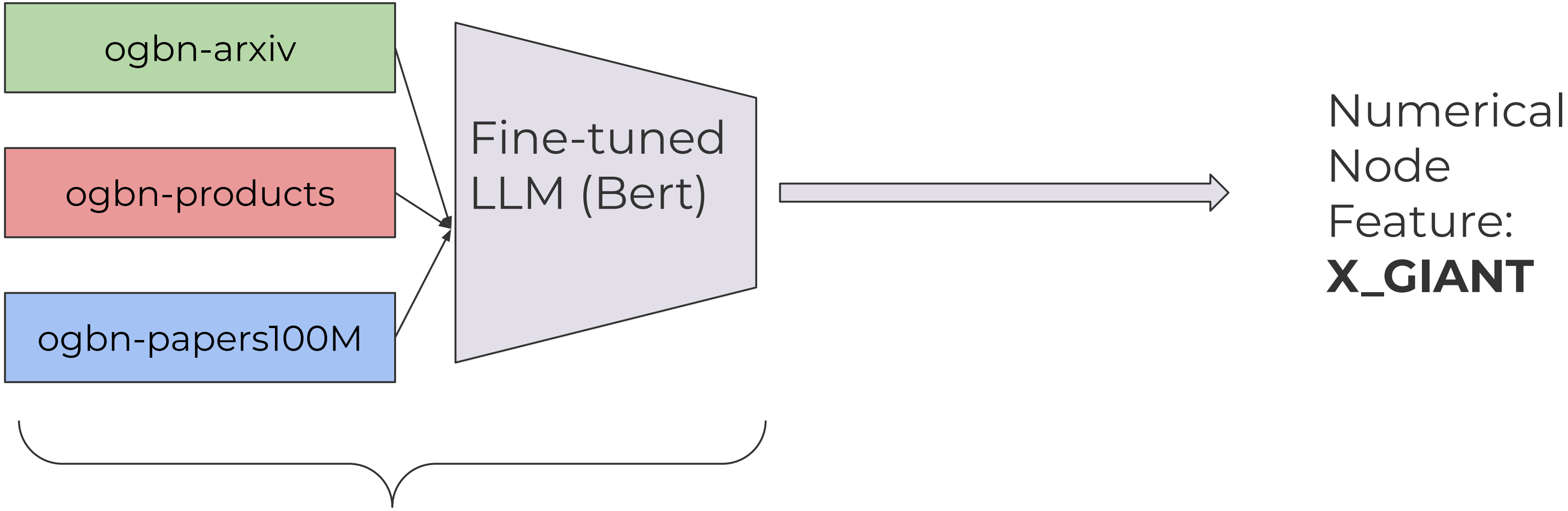
1. Dataset aspect:

   Trained on **one dataset at a time** (3 OGB dataset), rather than concatenate multiple datasets.

2. Prediction procedure

   Time consuming because utilized the XMC(Extreme Multi-label Classification) to **predict the similarity with each node** in the Adjacency Matrix (A).

Proposal



Node feature (text attribute) extraction

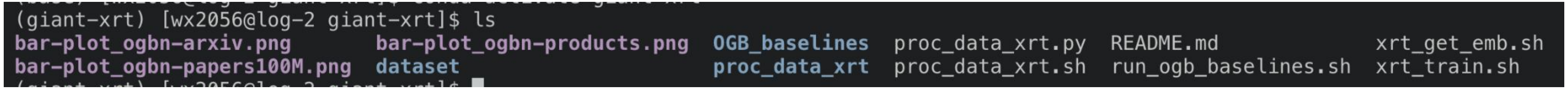
Environment Setup

Dataset Download: ogb packages

|---- params.json # hyper-parameters for GIANT-XRT pre-training
|---- X.all.txt # node raw text
|---- **X.all.xrt-emb.npy** # node embeddings from XR-Transformer >>>> for future contrastive learning
|---- xrt_models/ # XR-Transformer fine-tined models

NYU Greene HPC setup:

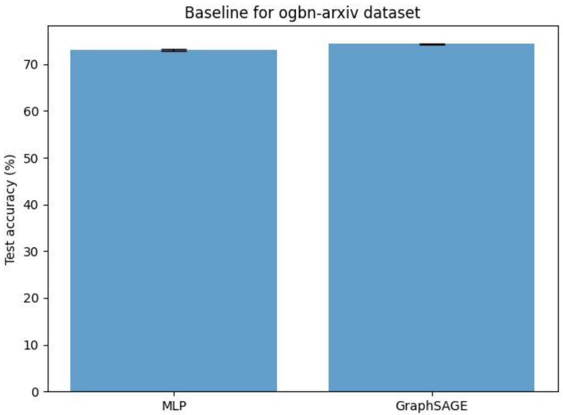
download giant-xrt pre-processed data under the ./proc_data_xrt folder

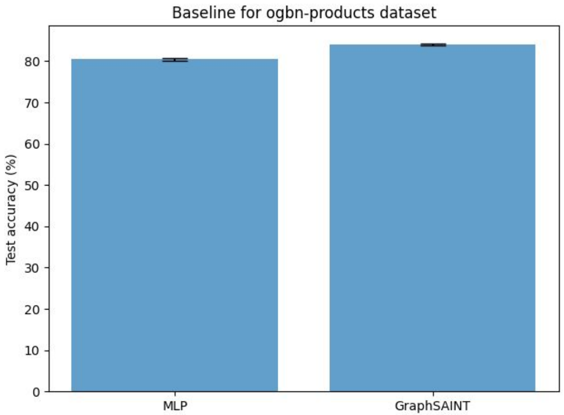


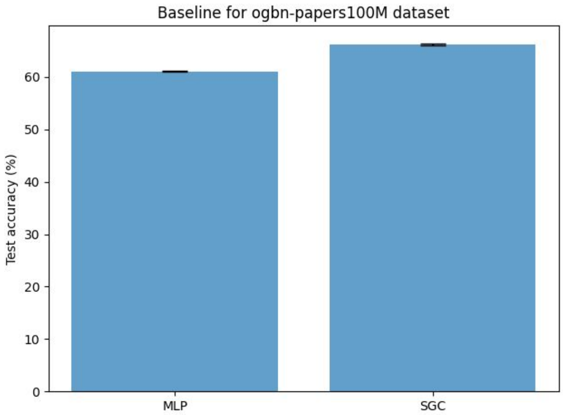
```
(base) [wx2056@log-2 giant-xrt]$ conda activate giant-xrt
(giant-xrt) [wx2056@log-2 giant-xrt]$ ls
bar-plot_ogbn-arxiv.png       bar-plot_ogbn-products.png  OGB_baselines  proc_data_xrt.py  README.md             xrt_get_emb.sh
bar-plot_ogbn-papers100M.png  dataset                     proc_data_xrt  proc_data_xrt.sh  run_ogb_baselines.sh  xrt_train.sh
(giant-xrt) [wx2056@log-2 giant-xrt]$
```

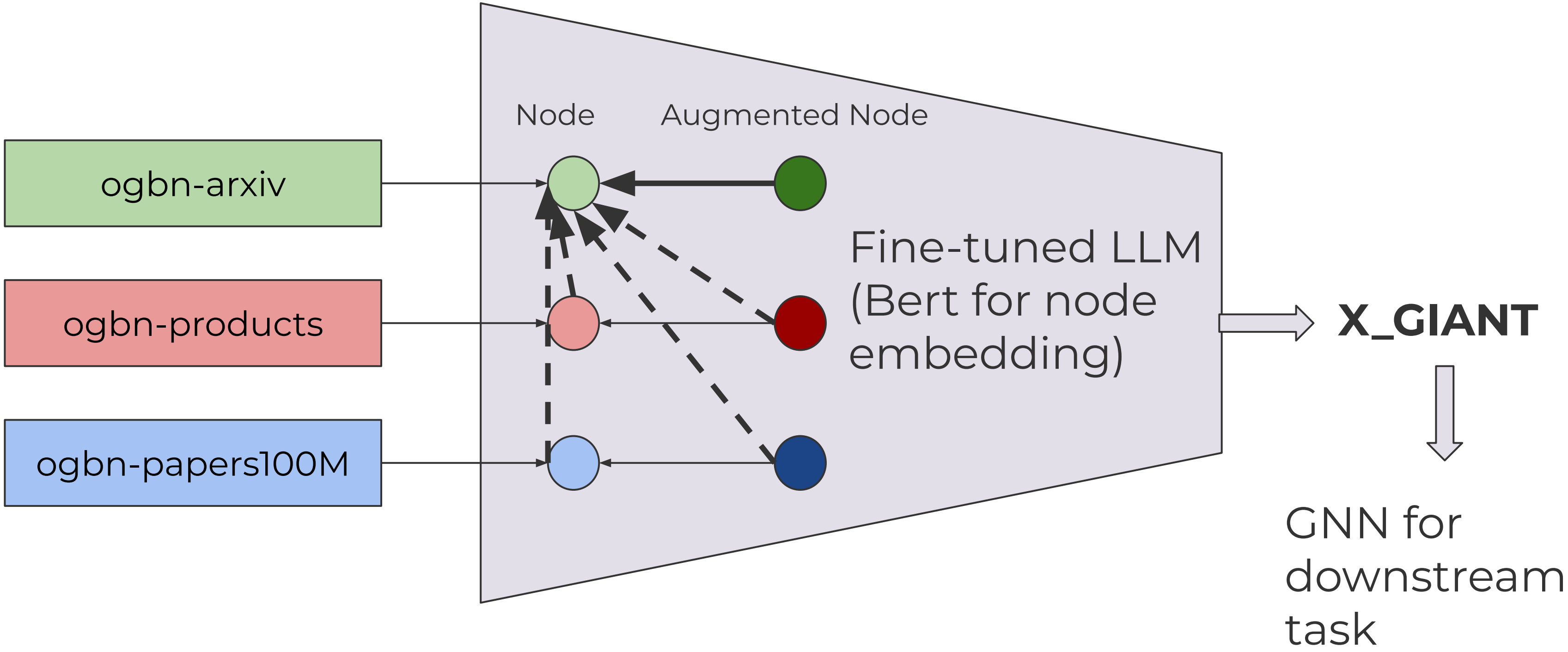
Baseline on Single dataset

```
dataset=***
gnn_algo=***
bash ./run_ogb_baselines.sh ${dataset} ${gnn_algo}
# for ogbn-arxiv: mlp/graph-sage
# for ogbn-products: mlp/graph-saint;
# for ogbn-papers 100M: mlp/sgc;
```


Baseline for ogbn-arxiv dataset


Baseline for ogbn-products dataset


Baseline for ogbn-papers100M dataset



ogbn-arxiv

ogbn-products

ogbn-papers100M

Node

Augmented Node

Fine-tuned LLM (Bert for node embedding)

X_GIANT

GNN for downstream task

Method

Dataset Concatenation

```
# Concatenate node raw text files
cat ./proc_data_xrt/ogbn-arxiv/X.all.txt \
        ./proc_data_xrt/ogbn-products/X.all.txt \
        ./proc_data_xrt/ogbn-papers100M/X.all.txt > ./proc_data_xrt/concatenated/X.all.txt
```

Hierarchical-XTransformer for XMC

Augmentation: 1) edge perturbation, add or remove edges with probability **P** 0.1 and 0.05 respectively

2) Add random Gaussian noise to node features

Contrastive Learning:

Within module xTransformer, update the training process from adjacency matrix label prediction to contrastively learning with pos samples and neg samples.

Reference

Citation:

- Chien, Eli et al. “Node Feature Extraction by Self-Supervised Multi-scale Neighborhood Prediction.” ArXiv abs/2111.00064 (2021): n. pag.