Final Report of GSoC

PyTorch Geometric Example Dataset using Pathway Commons and cBioPortal

Favour James, Final year BsC student at Obafemi Awolowo University
Google Summer of Code 2023 Contributor, NRNB

Overview of Project

- Abstract
- Introduction
- Project goals
- Project Methodology
- Data Preprocessing
- Data Integration
- Data Contribution
- Baseline Modelling
- GNN Modelling
- GAT Modelling
- Results
- Weekly Blog Posts

Abstract

Millions of people each year die from cancer worldwide. Understanding the underlying biological mechanisms that result in some individuals outliving others can help us produce better treatment strategies. Machine learning, specially neural network-based, models can help identify otherwise hard-to-identify patterns in patient data. Graph Neural Networks (GNNs) allow the development of predictive models based on known biological network data.

To aid in the development of such models, this project aims to generate an example dataset that integrates Pathway Commons (biological network) and cBioPortal (cancer patient) data for use with the popular PyTorch Geometric GNN library for the prediction of cancer patient overall survival. We provide scripts for processing similar datasets and example code to showcase how models can be generated. We hope this work will aid researchers in unlocking valuable insights into the role of biological pathways and genetic variation in cancer to help patients.

Introduction

Graph Neural Networks (GNNs) are a recent class of deep learning methods designed to perform inference on data described by graphs. Graphs can be used to model real-world phenomena such as transportation, social, and biological networks. Graphs are heterogeneously structured whereby the number of neighbors to each node is variable (as opposed to a fixed neighborhood size of images). GNNs are used to learn node embeddings through an aggregation of information from connected neighbors of a node through a process of message passing between nodes.

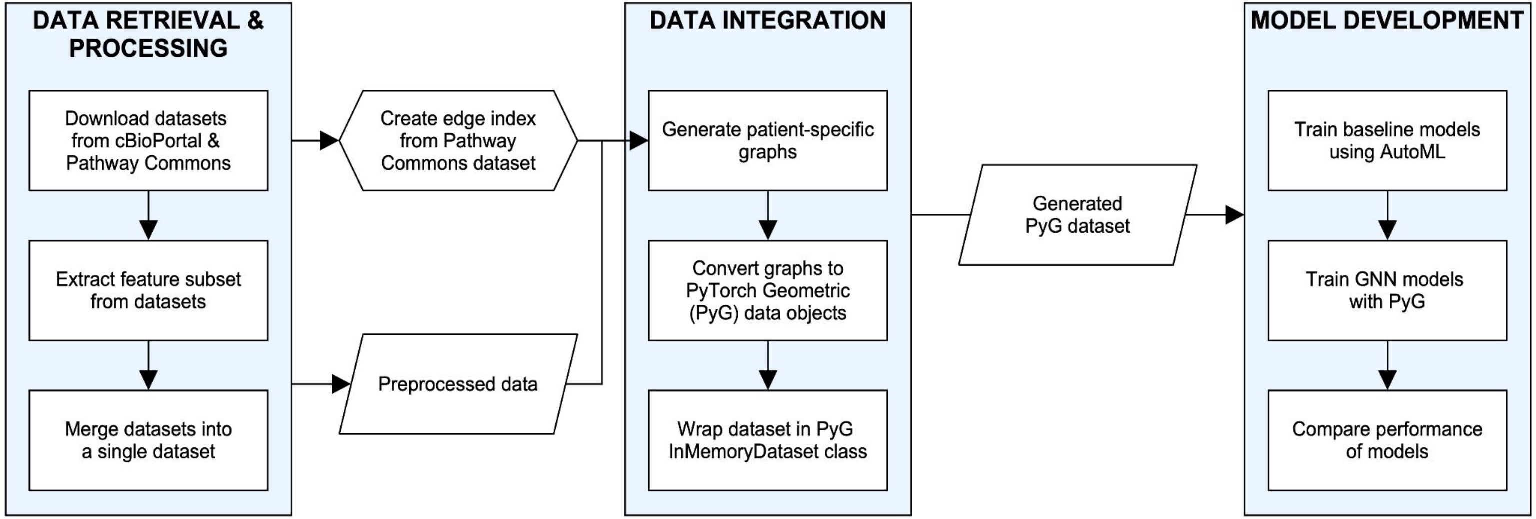
In this project, we use data from Pathway Commons, an aggregated database of millions of molecular interactions across 20 source databases. cBioPortal for Cancer Genomics is a database of multidimensional cancer genomics data collected from ~200 studies. Each patient from our cBioPortal dataset are represented by a graph where the nodes are genes and the edges represent the connection between genes. Input node values are provided by genomics data from cBioPortal while the edges are obtained from the Pathway Commons (using the Reactome subset) database.

Project Goals

- Create a dataset for use with PyTorch Geometric that includes cancer patient genomics and survival data from cBioPortal and network data to describe the graph structure for the genomics data from Pathway Commons.

- Develop example models using the developed dataset

Project Methodology



DATA RETRIEVAL & PROCESSING

Download datasets from cBioPortal & Pathway Commons

Extract feature subset from datasets

Merge datasets into a single dataset

Create edge index from Pathway Commons dataset

Preprocessed data

DATA INTEGRATION

Generate patient-specific graphs

Convert graphs to PyTorch Geometric (PyG) data objects

Wrap dataset in PyG InMemoryDataset class

Generated PyG dataset

MODEL DEVELOPMENT

Train baseline models using AutoML

Train GNN models with PyG

Compare performance of models

Data Preprocessing

First, the acc_tcga_2018(ACC: Adrenocortical Carcinoma) and brca_tcga_2018 (BRCA: Breast invasive carcinoma) datasets collected by were selected from cBioPortal representing small (N=78) and large (N=1082) collections of cancer patient data, respectively. Next, Gene expression features (N=9288) were extracted that overlapped with biological network data from Pathway Commons. Additionally, overall survival time in months was extracted as the value to be predicted. The dataset gotten from these steps were uploaded to Zenodo.

The following notebooks show these steps:

- acc_tcga - acc_tcga_preprocessing
- brca_tcga - brca_tcga_preprocessing

Results gotten:

1. graph_idx: This consists the gene features of the patients.

2. graph_labels: This consists of the overall survival(months) of each patient.

3. edge_index: This consists of the total edges.

Data Integration

In this stage, an edge index (N=271771 edges) was generated using Pathway Commons v12 data in a tabular format. To convert to a PyG Dataset, a list of graphs were created from the preprocessed dataset first. Then, these two data types were integrated which resulted in patient-specific graphs which were then converted into PyG data objects. Finally, these graphs are wrapped using the InMemoryDataset class for use with PyG. The dataset gotten from these steps were uploaded to Zenodo, a general-purpose data repository meant for long-term storage of academic datasets.

These steps are shown in the provided notebooks:

- acc_tcga: acc_tcga_data_integration
- brca_tcga: brca_tcga_data_integration

The Dataset statistics is shown here: Dataset Statistics

Data Contribution

For dataset contribution, the breast_cancer dataset was selected. The choice of the breast_cancer dataset was motivated by its larger size in comparison to the acc_data. To ensure compatibility with other datasets in the PyG GitHub repository, the notebook for integration was transformed into a Python script. Subsequently, this script was formatted to successfully clear all essential tests before it was deemed ready for contribution to the PyG repository. As the final step, a pull request was initiated and is presently undergoing review within the PyG community.

This python script can be accessed here: brca_tcga_python_script
Pull Request: pull_request_for_contribution

Another repository was created which contains only the notebooks and python script for the brca_tcga dataset for users only interested in this dataset.
This repo can be accessed here : dataset_repository

Baseline Modelling

Baseline Models for the acc_tcga and brca_tcga were developed for comparison with Graph Neural Network(GNN) models using the Automated Machine Learning (AutoML) library: FLAML Fast Library for Automated Machine Learning (FLAML).

The steps used are shown in the following notebooks:

- acc_tcga: acc_tcga_baseline_model
- brca_tcga: brca_tcga_baseline_model

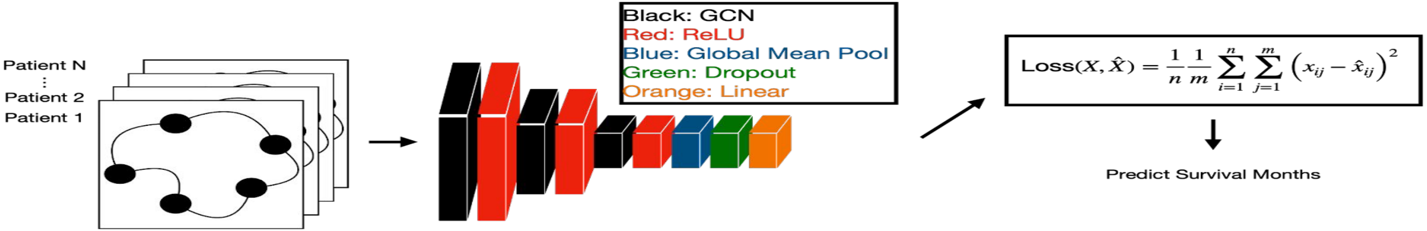
GNN Modelling

Graph-based models were developed for each cancer type. GNN models for each cancer type (i.e., ACC and BRCA) are developed employing Graph Convolutional Networks (GCNs). GCNs are class of deep learning architectures tailored for processing graph-structured data. They effectively learn patterns and relationships within the intricate networks of biological data.

The steps used are shown in the following notebooks:

- acc_tcga: acc_gnn_model
- brca_tcga: brca_gnn_model

The model architecture is shown below:

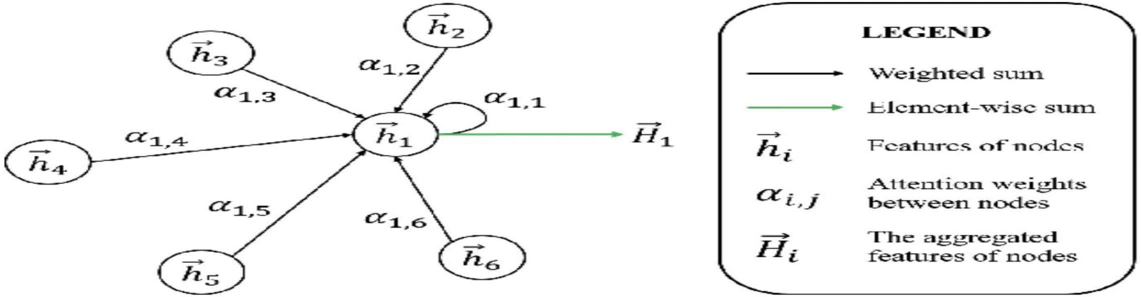


GAT Modelling

Graph Attention Networks was also used for modelling with the BRCA dataset. GATConv, a part of GAT, performed better than Graph Convolutional Networks (GCNs). This is because GAT pays more attention to important neighbors, capturing data relationships better.
The steps used are shown in the following notebook:

- Brca_tcga_gat_model

The model architecture is shown below:



Results

Preliminary results demonstrate that the GNN model performs similarly to the FLAML-based model for the smaller ACC model in predicting overall survival time. In contrast, the GAT model for Breast Invasive Carcinoma (BRCA) outperforms both the BRCA FLAML and Graph Convolutional Network (GCN) models. We believe this observation between the models is based on the larger BRCA dataset.

Given the small size of the ACC data, it was split to only train and test sets. However, for the BRCA data, a validation split was included. During the development of the baseline models, only train and test splits were used. This was done to fit with the requirements of FLAML.

Modelling statistics can be accessed here: modelling_statistics

Blog Posts

During this project, I was writing weekly blog posts on my medium to document my progress each week. Below are the links to the reports from week 1 to week 10. At the time of the final report, week 11 and week 12 blog posts had not been published.

- Community Bonding period and week 1
- Week 2
- Week 3
- Week 4
- Week 5
- Week 6
- Week 8
- Week 9
- Week 10

During this project too, a research poster was created which was presented at the 2023 Data Science Nigeria-ai bootcamp, which won the best poster award. This poster would also be presented at the Deep Learning Indaba,2023 happening at Accra, Ghana.

The poster can be accessed here: poster link