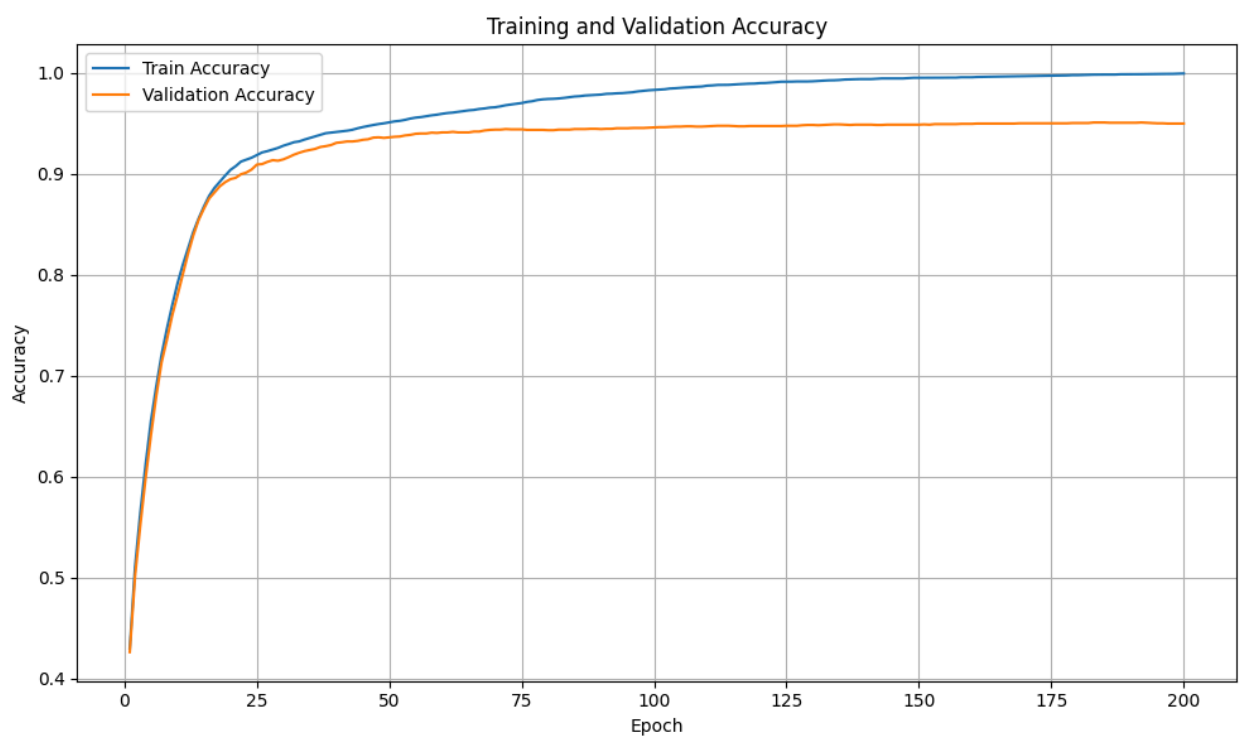
GNN Project Report

By Ido Beigelman

For this project, I utilized the Coauthor CS dataset provided by the Deep Graph Library (DGL). In this graph-based dataset, each node represents an author, and an edge exists between two nodes if the corresponding authors have co-authored at least one research paper. The node features are derived from a bag-of-words representation of keywords extracted from the authors' published papers. Each node is labeled according to the author's primary field of study, serving as the classification target in our experiments.

The Coauthor CS dataset contains 18,333 nodes and 163,788 edges, representing authors and their co-authorship relationships, respectively. Each node is associated with a 6,805-dimensional feature vector, constructed using a bag-of-words representation of keywords from the author's publications. The dataset includes 15 distinct classes, corresponding to different research fields within computer science.

We began our experiments with the GraphSAGE model, exploring various configurations involving different hidden layer sizes and numbers of layers. To determine the optimal number of training epochs, we evaluated model performance across multiple runs. Our results showed that the model's accuracy converged by epoch 200. An example plot of the training and validation accuracy over epochs is provided below:



Based on our observations regarding the optimal number of training epochs, we proceeded to train and evaluate various configurations of hidden layer counts and hidden layer sizes. The validation accuracy for these different hyperparameter settings is shown below:

Hidden Layer Size	Number of hidden layers	Validation Accuracy
16	1	0.9509
16	2	0.9425
16	3	0.9356
32	1	0.9479
32	2	0.9457
32	3	0.9397
64	**1**	**0.9534**
64	2	0.9474
64	3	0.9422
128	1	0.9528
128	2	0.9468
128	3	0.9436
256	1	0.9525
256	2	0.9487
256	3	0.9438

The model achieved its best performance with a single hidden layer of size 64. Overall, we observed a decline in accuracy as the number of hidden layers increased. This suggests that an author's immediate co-authors provide the most relevant information for predicting their field of study. Incorporating higher-order neighborhood information—such as co-authors of co-authors—appears to introduce noise rather than improve performance.

Building on the best-performing configuration, we attempted to further improve the model by experimenting with different aggregator types. While the default aggregator in GraphSAGE is the mean aggregator, we also evaluated sum, max, and LSTM-based aggregation. However, these variations did not result in any noticeable performance differences. All aggregator types yielded comparable results, indicating that the choice of aggregator had minimal impact in this specific setting. The results for each aggregator type are summarized below:

Aggregator	Validation Accuracy
Mean	0.9534
Sum	0.9534
Max	0.9534
LSTM	0.9534

To further improve model performance, we incorporated attention mechanisms by implementing a Graph Attention Network (GAT). As in previous experiments, we explored various hidden layer sizes and number of layers to identify the optimal architecture. Additionally, we introduced the number of attention heads as an additional hyperparameter to tune. The tables below present the average validation accuracy across different configurations of these hyperparameters.

Hidden Layer Size	Average Validation Accuracy
16	0.9181
32	0.923
64	0.9263
128	0.9305
256	0.9338

Attention Heads	Average Validation Accuracy
1	0.9133
2	0.9223
4	0.9284
8	0.9323
16	0.9358

Number of Layers	Average Validation Accuracy
1	0.9275
2	0.9275
3	0.9235

To conclude, although the attention-based network (GAT) is architecturally more complex than GraphSAGE, it did not lead to improved performance in our experiments. Ultimately, the best-performing model was the GraphSAGE network with a single hidden layer of size 64. When evaluated on the test set, this configuration achieved a final accuracy of 94.74%.