CSGS Hackathon - Binary Bandits

Dominik Soós
David Calano

October 31, 2024

For this assignment, I watched the following videos:

1 Introduction

The classification of drug interactions comes in two primary varieties, Drug-Target Interaction (DTI) and Drug-Drug Interaction (DDI). We went with Drug-Drug Interaction for a challenge working with multi-type classification and predicting the effects that drugs have on either other and their potency.

2 Data

One of the challenges was to get our hands on data. DrugBank is the most popular dataset but the full dataset requires a subscription. We were able to find the Theraputic Data Commons [1] library which provides free access to a recent version of the DrugBank dataset and helper functions for working with the data.

2.1 DrugBank

The most popular dataset for Drug-Drug Interaction (DDI) has been used extensively to test the performance of multi-prediction tasks. There are 1706 unique drugs in this dataset with 191,808 drug-drug interaction pair data points. We were able to use DrugBank through TDC and then parse the data through DeepChem to obtain molecular data for our graph neural network. The DrugBank dataset is very imbalanced with two classes dominating the distribution.

Figure 1 illustrates the class imbalance. We can observe the difference in the distribution after we performed the data augmentation. We simply kept doubling the number of samples until a desired number of samples were reached for each class.

[1] `https://tdcommons.ai/`

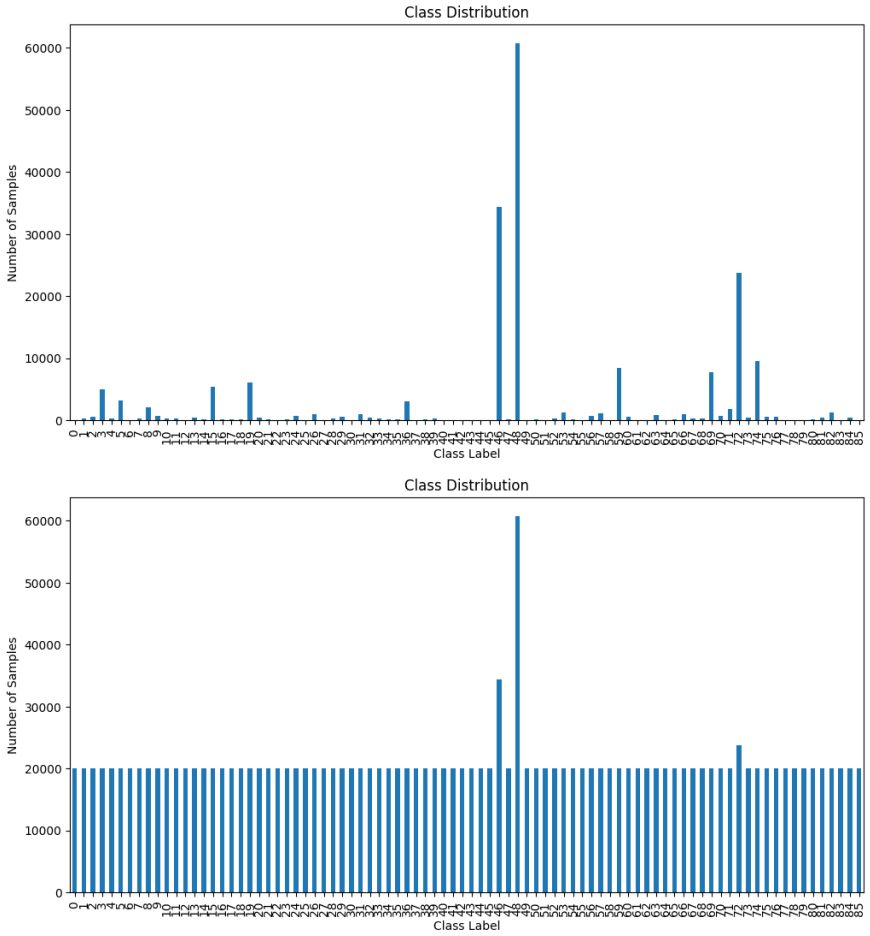

Figure 1: DrugBank dataset before (top) and after (bottom) data augmentation

3 Methodology

We implemented a graph convolutional network since the data itself is in a graph format.

3.1 Model Architecture

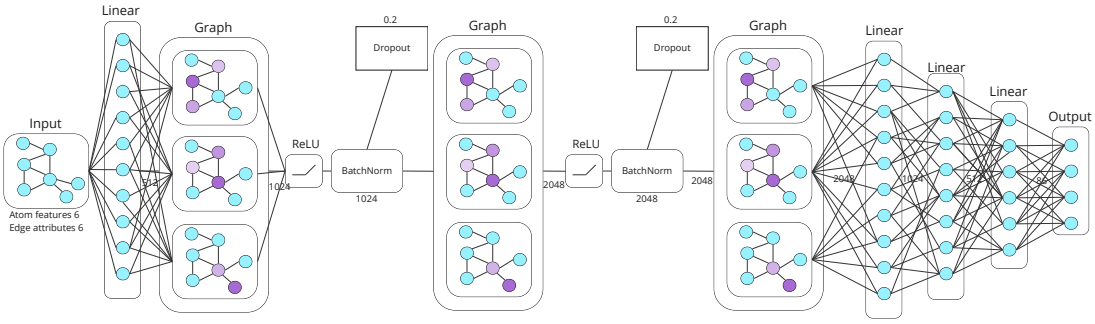

Figure 2: Model architecture

3.2 DrugBank

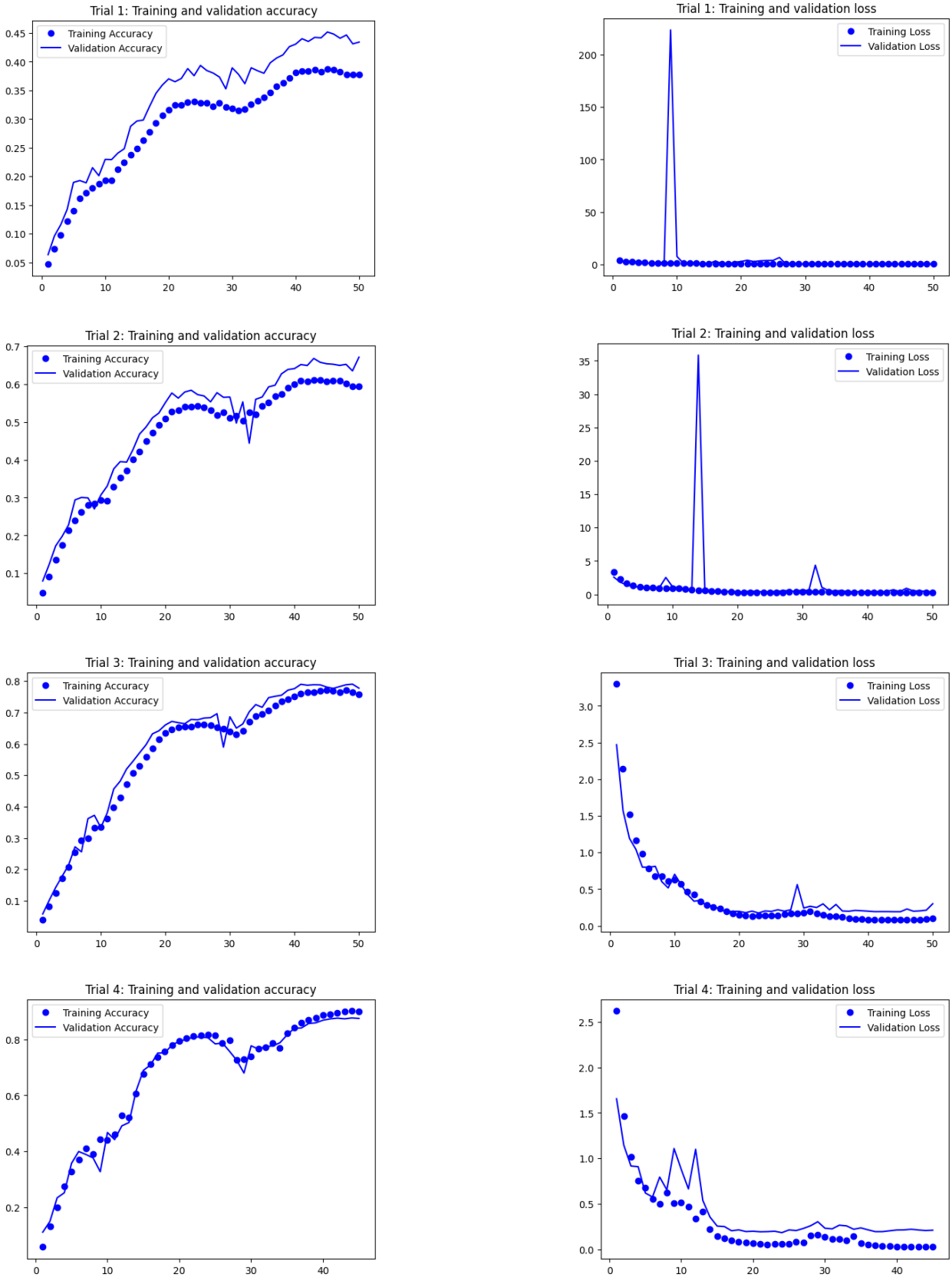

Figure 3: Training accuracy and loss data from DrugBank trials

4 References