CS 598 DLH Project Proposal - Team 31

Avinash Baldeo abaldeo2@illinois.edu
Jinfeng Wu jinfeng4@illinois.edu
Hao Zhang haoz8@illinois.edu

---

1. General problem

The paper we have selected to reproduce is "CNN-DDI: a learning-based method for predicting drug–drug interactions using convolution neural networks" [1]. This research relates to the issue of drug-drug interactions (DDIs) in pharmaceuticals development. Antagonistic DDIs are reactions between two or more drugs that may lead to adverse effects that diminish the efficacy of the drugs involved. Since these drugs are expensive to develop it is important to be able to predict DDIs based on properties of drugs. Knowing if two drugs interact is also useful since drugs similar to either of the two are more likely to interact and cause the same effect [1]. The paper proposes a novel method, CNN-DDI, which utilizes convolutional neural networks (CNNs) to predict DDIs by learning from a chosen combination drug features such as categories, targets, pathways, and enzymes. It builds on a previous work “A multimodal deep learning framework for predicting drug-drug interaction events” [2], which uses a DNN model along with four drug features (Target, Enzyme, Pathways and Substructure) to predict DDIs. In the CNN-DDI model, a feature selection framework is constructed to select the best combination of drug features, which is stated to be the Target, Enzyme, Pathways and Category.

2. Approach

To reproduce the results of the paper, the following steps will be followed:

1. Collect dataset from data source(s). We will use data collected in DDIMDL repo and add category data from DrugBank.
2. Do feature extraction to construct feature vectors.
3. Perform similarity calculation and create drug similarity matrices using Jaccard similarity (will attempt Cosine and Gausine similarity if time permits, but for now out of scope)
4. Implement the CNN-DDI model as described in the paper, including the convolutional layers, residual block, and activation functions.
5. Train the CNN-DDI model using the dataset collected and hyperparameter settings found from the CNN-DDI & DDIMDL papers.
6. Evaluate the result using following metrics from the paper:Accuracy, F1-score, micro-averaged AUPR and micro-averaged AUC
7. Compare our results with those reported in the CNN-DDI paper.

3. Data Access

The primary dataset used by CNN-DDI is from the DDIMDL Github repository (https://github.com/YifanDengWHU/DDIMDL). The DDIMDL paper classifies DDIs' events into 65 types and includes 572 drugs with more than 70,000 associated events. The data was originally collected from the DrugBank website (https://go.drugbank.com/) using a web scraper and then processed and stored into a SQLite database (event.db). To utilize this dataset for CNN-DDI, we need to extract the 1622 category types for the drugs in the database from the Drugbank and store it.

4. Tested Hypotheses

The primary hypothesis that will be tested is that the CNN-DDI model, which utilizes a feature selection framework and a novel CNN architecture, can accurately predict drug-drug interactions and outperform other the models mentioned in the paper (Random forest, Logistic Regression, K-nearest neighbor, Gradient boosted Decision Tree, & DDIMDL). This hypothesis will be tested by implementing the CNN-DDI model according to the approach mentioned in the paper and training on the collected dataset with the same/inferred settings. Afterwards, we will compare the results with table 3 and 4 from the paper to our results.

5. Ablations

To understand the contribution of different components to the model's performance, we will attempt to do both feature and model ablation study.
For feature ablation, we wil evaluate the model's performance by individually removing drug categories, targets, pathways, and enzyme features. This is done in the paper and will allow us to verify the claim made that the drug category is an effective predictor for DDIs.
For model ablation, we plan to assess the impact of removing the residual block on the model performance and try different loss functions such as Binary Cross-Entropy Loss.

6. Computation Feasibility

For each convolutional layer:
Parameters: (filter_height × filter_width × input_channels × output_channels) + output_chanels
Operations per input: (assuming stride 1)
(output_height × output_width) × (filter_height × filter_width × input_channels) × output_channels

For each fully connected layer:
Parameters:
(input_features×output_features)+output_features
Operations:
2×input_features×output_features

Based on calculations using these formulas, the CNN-DDI model has approximately 149.89 million parameters and requires about 731.38 million operations for a single forward pass.
This would require a modern GPU with several GBs of memory. Training the model (backwards pass) would require 2-3 times more.
Google Colab offers T4 GPU with 16 GBs memory. Colab Pro offers A100 and V100, limited to 100 compute units for $14 per month. We plan to run one epoch on the Colab T4 GPU and see how many compute units it costs before determining if additional computational resources offered in Colab Pro are required for the full 100 iterations of training required. As of now, we believe that this is computational feasible for us to do.

7. Existing Code Reuse

Since the source code for CNN-DDI is not publicly available, our plan for reproducing is to reuse as much of the existing code from the DDIMDL Github repo as possible. Our first task is to extract drug Category data from the Drugbank, which is not included in the existing code. We will also need to modify the feature construction code to include this as part of the input.
CNN-DDI constructs a similarity matrix from the input features, in the same way that DDIMDL does it, so we should not need new code for this. The next task will be to implement the CNN model for predicting DDIs from scratch architecture according to the paper's design. We will have to implement the training ourselves. For evaluating and doing the cross-validation of the results, we will retrofit the existing code.

8. References

[1] Zhang, C., Lu, Y. & Zang, T. CNN-DDI: a learning-based method for predicting drug−drug interactions using convolution neural networks. *BMC Bioinformatics* 23 (Suppl 1), 88 (2022). https://doi.org/10.1186/s12859-022-04612-2

[2] Deng Y, Xu X, Qiu Y, Xia J, Liu S. A multimodal deep learning framework for predicting drug-drug interaction events.
In: 2020 15th IEEE international conference on automatic face and gesture. 2020