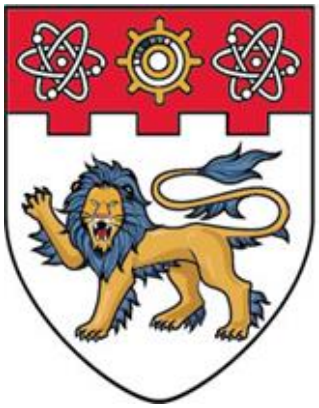

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

BS6207 ADVANCED ARTIFICIAL INTELLIGENCE FOR BIOMEDICAL DATA SCIENCE

Final Project

Proteins and Ligands Binding Classification

NIYUXIN

11/05/2021

Contents

1. Problem definition

The aim of this project is to establish a deep learning model fed coordinates and atom types of protein and ligand to predict whether they bind with each other or not. And export a file contains the top 10 ligands with highest probability to bind with specific protein.

2. Preprocessing

2.1 Feature Engineering

6D dimension matrix will be used to represent ligands and proteins, each row vector means one atom, a 3 out of 6 dimension means the coordinates of the position of atoms, and the rest of the features were used to describe the atoms:

2 digits of one-hot encoding to represent the atom type: 1,0 means polar and 0,1 means hydrophobic.

1 digit to represent the type of molecule: 1 means protein and -1 means ligand.

The features of ligand '0003' was shown in Fig 1. There are 6 atoms to represent this ligand, and the first three columns shows the coordinates of each atoms, the rest 3 columns describe the type of the atom and ligand.

```
lig_data_set["0003"]

array([[ 3.308, 10.315, 25.119,  0.  ,  1.  , -1.  ],
       [ 5.761,  7.669, 23.663,  1.  ,  0.  , -1.  ],
       [ 6.614,  3.592, 20.328,  0.  ,  1.  , -1.  ],
       [ 9.714,  1.305, 18.22 ,  0.  ,  1.  , -1.  ],
       [ 9.77 ,  7.655, 26.167,  0.  ,  1.  , -1.  ],
       [14.336,  7.5  , 26.728,  1.  ,  0.  , -1.  ]])
```

Fig 1 '0003' ligand representation with atoms

2.2 3D Grid Generation

In the dataset, proteins are much larger than ligand, then I cropped the complex to a defined size of X-A°(default 20,block size, 2*max distance) cubic box shows in Fig 2 that focused at the geometric center of a ligand followed by [1]. In this case, geometric information out of the cubic will be eliminated. We then discretized the positions of heavy atoms using a 3D grid with Y-A°(default 1) resolution(R) and move the precise geometric coordinates into its nearest node of the cubic. In this case, smaller resolution should make the node more precise compared to the original geometric coordinates, and max distance is the absolute distance between the cubic face and the center, larger max distance will contain more information about the geometric of the proteins and ligand. The equation of block size that will be input into CNN model shows below.

$$Block\ size = (2 \times \max distance)/resolution$$

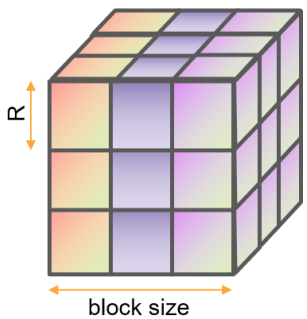

Fig 2 grid set

2.3 Negative Sampling and Dataset Splitting

I sample all the possible tuple (protein, ligand) and as long as the name of protein and ligand in the

sample tuple were not same, they are negative samples. Size of Total dataset was 9,000,000 and size of negative sample dataset was 8,997,000.

Dataset was split into train dataset and validation dataset with the ratio of 4:1.Table 1 shows the details of the distribution of the dataset.

Table 1 Dataset distribution

	Positive Data	Negative Data
Train Dataset	2,400	7,197,600
Validation Dataset	600	1,799,400
Total	3,000	8,997,000

3.3.1 Training and validating method

The size of negative data is much larger than that of positive dataset. Since the purpose of this project is to predict positive binding. In training stage, there is no need to use all the negative data to let the model focus more on the negative data. In this case, the negative data in training stage was extracted from the total negative dataset and the size of it is totally equal to the size of positive training data set.

The validation stage is also the same, I randomly extracted negative data in the validation negative dataset with the same size (600) of the validation positive dataset.

3.3.2 Validation dataset to evaluate final test performance

The label of test data is not known and we need to export the top10 ligands that have large percentage to bind with protein. I generate the data set based on validation dataset to experiment, mimic and evaluate the test data performance.

As Table 2 the shows, I use the total number of proteins in validation dataset. Since their 824 different combination for each protein in test data. 823 negative samples were randomly sampled in validation negative dataset. And top 10 ligands with highest probability that model predict will also be checked to see whether the corresponding ligand is in the top 10 list or not.

Table 2 Validation imitation for test

	Valid Data	test Data
Protein data	600	824
Possible combination	600 X 824	824 X 824
Total	494,400	678,976

3.4 Metrics

3.4.1 Metrics for training and validation stage

In the training and validation stage, accuracy is used ,the equation of accuracy shows in below,

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where TP, TN,FP,FN means true positive cases, true negative cases, false positive cases and false negative cases, respectively.

3.4.2 Metrics for test experimental stage

I also define 'success rate' SR in this stage, one shot success means the positive ligand shows in the top 10 ligands of one protein predicted by model, then the equation can be established below,

$$SR = \frac{n_s}{N} \times 100\%$$

Where n_s means the number of one ligand shows in the top 10. N is the total number of proteins, which

here is 600.

The ranking position of the correct ligand in the total top 10 ligands list also matters, then nDCG(Normalized Discounted Cumulative Gain) in Recommendation System Region can also be used. The Equation of nDCG shows below,

$$nDCG_p = \frac{DCG_p}{IDCG_p}$$

Where p is the length of ranking list, DCG is the discounted cumulative gain and IDCG is the ideal discounted cumulative gain. The formulation of DCG and IDCG is same instead of that ideal DCG is the best and largest DCG of the ranking list. The formulation shows below,

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log_2(i+1)}$$

Where rel_i means the relation between the i-th ligand and the target protein, 1 means that the ligand binds with protein and 0 means ligand does not bind with protein. So, the ideal case $IDCG_{10} = \frac{1}{log_2 2} = 1$.

3. Model Establishment

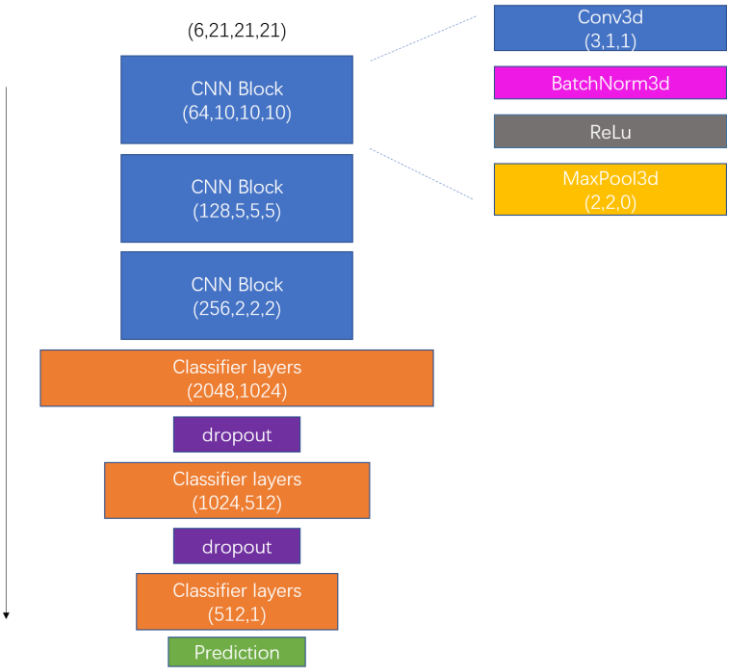

Fig 3 NN model

The basic deep neural network architecture of this project shows here in Fig 3, this model contains three 3D CNN blocks, 3 classifier layers with 2 dropout layers. The dimension of input of the model (6,21,21,21), where 6 means the sum of the three features between protein and ligand and the rest of dimension means the coordinates. (64,10,10,10),(128,5,5,5),(256,2,2,2) are the output dimension of these three CNN blocks respectively. The setting of convolutional layer and max pooling layer in CNN block shows in Table 3.

Table 3 details of layer in CNN block

	Kernel size	stride	padding
Conv3d	3	1	1
MaxPooling3d	2	2	0

4. Training procedure

After the structure of the deep learning neural network was done, some hyper parameters of the model could be tuned by selecting different parameters and evaluate the validation performance.

4.1 Random Rotation of training data

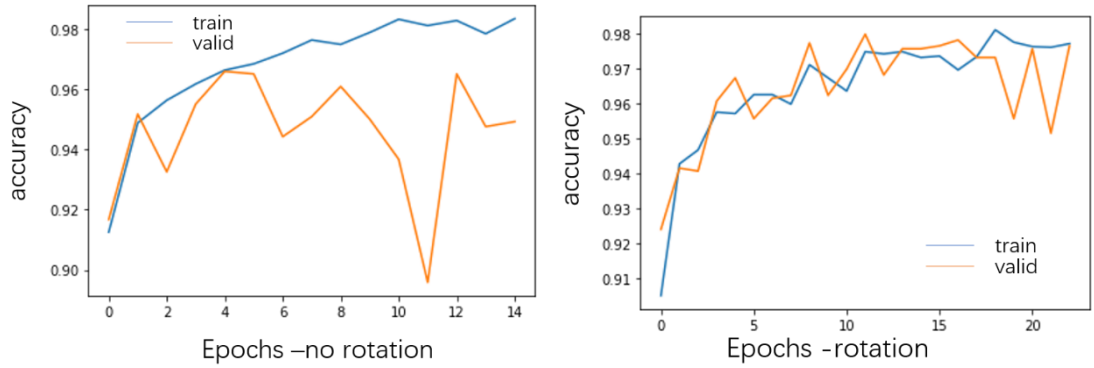

Fig 4 performance of rotation in training data (left: no rotation; right: rotation

The result shows in Fig 4 shows that using random rotation in training data can make a better train for the model and the accuracy of model in validation data will increase significantly.

4.2 dropout parameter selection

The first hyper parameters that were tuned is the dropout probability in the both dropout layers between two classifier layers. 0.2,0.4,0.6,0.8 were selected as the values of dropout probability.

In Fig 5, it can be seen that 0.8 probability can reach the highest accuracy and lowest loss, and the best validation accuracy of those 4 dropout probability are 0.98,0.98,0.9767,0.9825, respectively, then 0.8 dropout probability was chosen for further hyper parameters tuning.

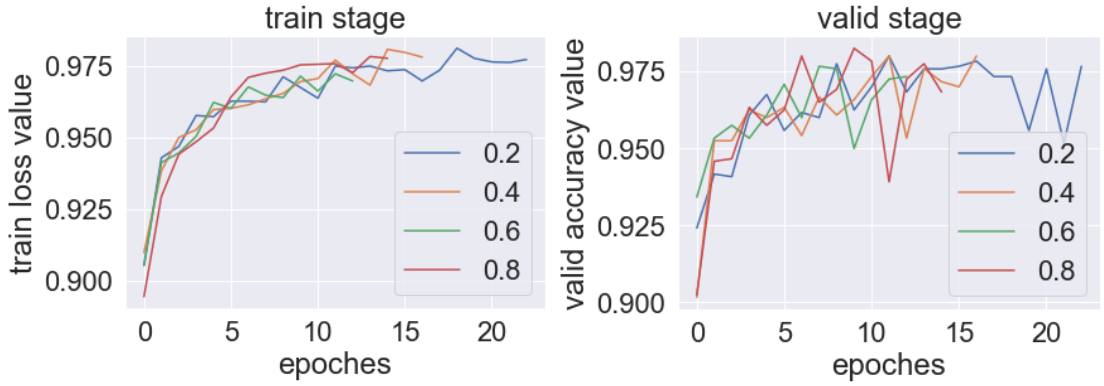

Fig 5 accuracy distribution

4.3 learning rate selection

The second hyper parameters that were tuned is the learning rate of the Adam optimizer.0.1,0.01,0.001,0.0001 were selected as the values of learning rate.

In Fig 6, it can be seen that the performance of 0.1 will make the mode hard to fit the data, the accuracy of 0.1 learning rate is only approximately 55%. While performance of other learning rate is good and accuracy of 0.001 is the highest. In this case the learning rate 0.001 was chosen.

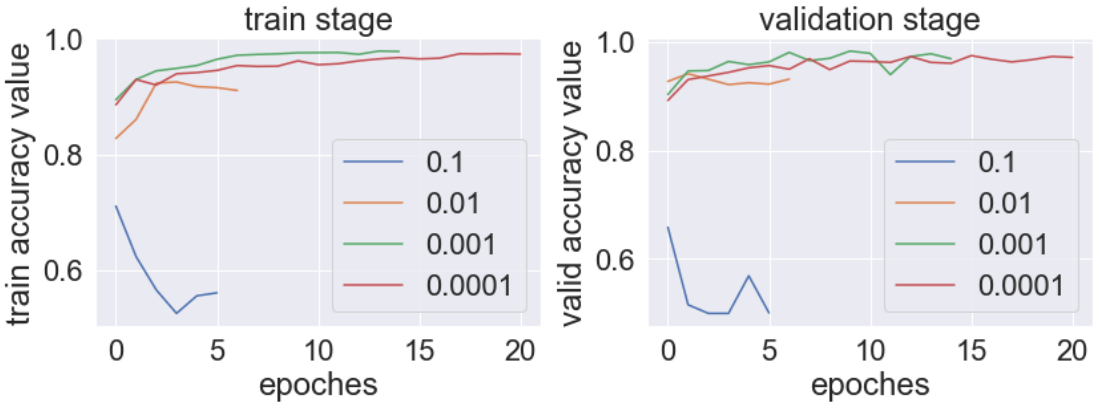

Fig 6 accuracy distribution of learning rate tuning

4.4 maximum distance selection

Maximum distance is an important parameter of the preprocessing stage. In the same resolution, large Maximum distance compared to center will contain more geometric information of protein and ligand. In this case, 2 different size of the block were chosen. The result shows that larger information does not give the better performance. Then 20 is chosen.

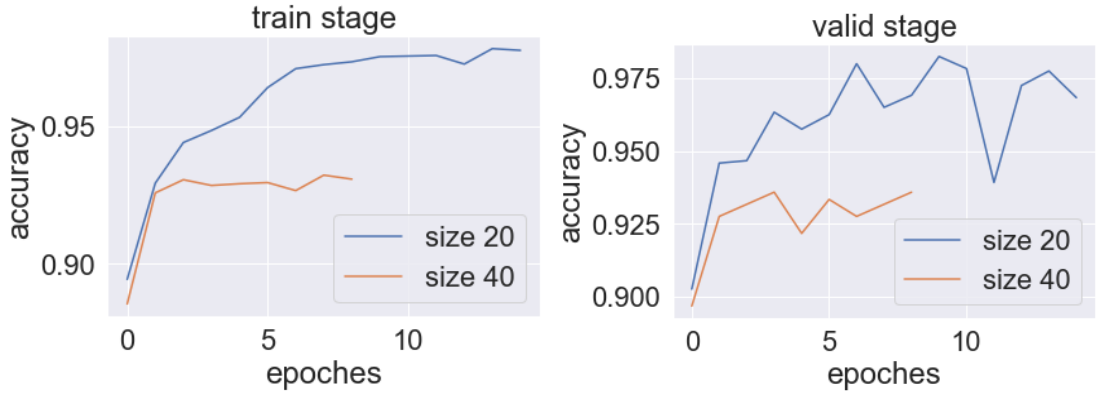

Fig 7 accuracy distribution of max distance

4.5 resolution selection

Resolution shows the degree of precision, smaller resolution makes the cubic denser, but the information in the cubic does not change. The result in Fig 8 shows the performance of 1 and 0.5 do not have severe difference, and since model with 1 resolution can train more faster, then 1 was chosen.

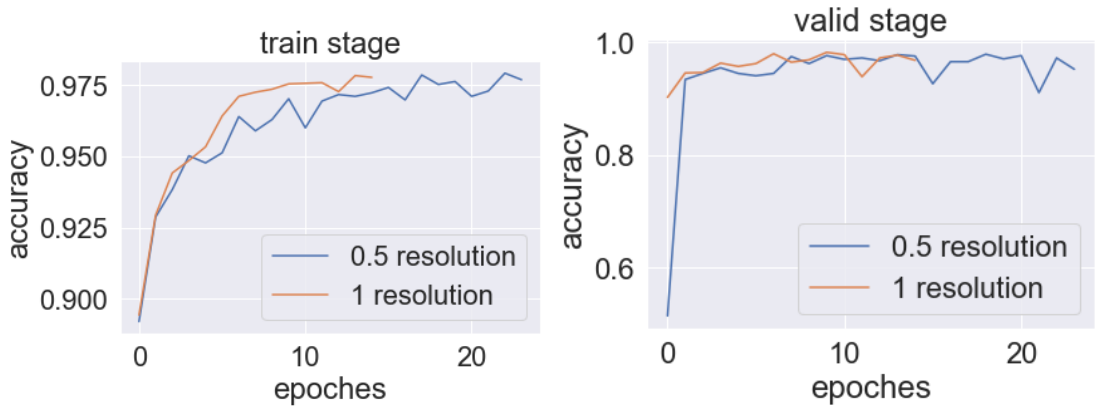

Fig 8 accuracy distribution of resolution

5. Experimental study

The performance of imitation test by using valid dataset shows in Fig 9 and Fig 10. The result shows that this model is very good at classifying true positive data (high recall), but not good at false positive data(low precision).

	precision	recall	f1-score	support
0.0	1.00	0.97	0.98	493800
1.0	0.03	0.99	0.07	600
accuracy			0.97	494400
macro avg	0.52	0.98	0.52	494400
weighted avg	1.00	0.97	0.98	494400

Fig 9 classification report

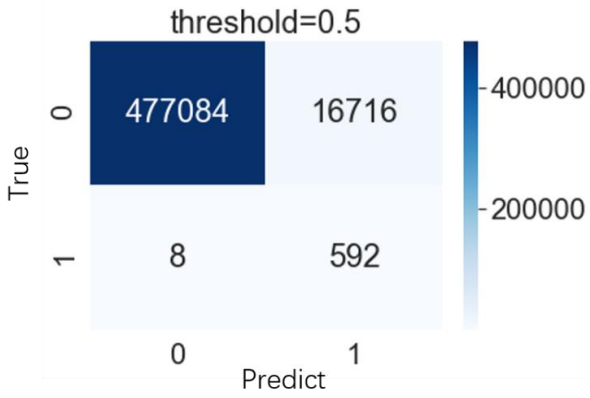

Fig 10 Confusion matrix of imitation test

In this case, ranking evaluation and shot accuracy are very important to evaluate the model's performance. The success rate is 0.953 and nDCG is 0.776, which means in total 600 proteins, there are 95.3% accuracy that the target ligand exists in the top 10 list, and the average ranking of the target ligand in the top 10 is less than 1st (nDCG=1) but larger than 2nd (nDCG=0.631). In this case, the model is very good.

6. Summary

This model could reach 95.3% accuracy in success rate and 0.776 of nDCG. And the average ranking of the target ligand in the top 10 list in a specific protein is between 1st and 2nd. The performance in test dataset will be expected in the similar performance.

7. Reference

[1]. Stepniewska-Dziubinska, Marta M., Piotr Zielenkiewicz, and Pawel Siedlecki. "Development and evaluation of a deep learning model for protein–ligand binding affinity prediction." Bioinformatics 34.21 (2018): 3666-3674.