Computational Biology: Homework 3

Instructed by *Jianyang Zeng*

Due on 24 Mar,2020

Xinyi Zhao YaoClass 81 2018013417

In this homework, we want to predict protentíal RBP binding sites in SARS-COV-2. I preprocess the data mainly following the paper [1]. And I do some experiments to check whether employ some other methods could increase the prediction accuracy(roc-auc score). And then we use the prediction model to predict the potential RBP binding sites.

1 Model

1.1 Preprocess data

1.1.1 Paper method

We use the basic bag-of-words model. We fix the length k and count how many times it appear in both target region and the position around the target regions(around 300). And referred to the paper, I use length k as 1,2,3. And use the stride 2 to get the string of length k=4,5,6 and then use some special features like length to record the information. In total, I use 271 features to represent one RNA. We define these 271 features to be the base features.

1.2 Doc2vec method

Moreover, similar to homework 2, I also use the doc2vec to represent the downstream and upstream part and the target part as shown in fig 1. And here, we use over 10w data in different protein data to train the representing model in order to increase the accuracy of the model.

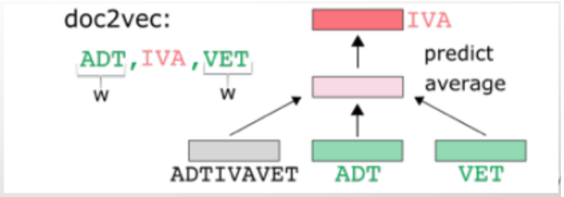
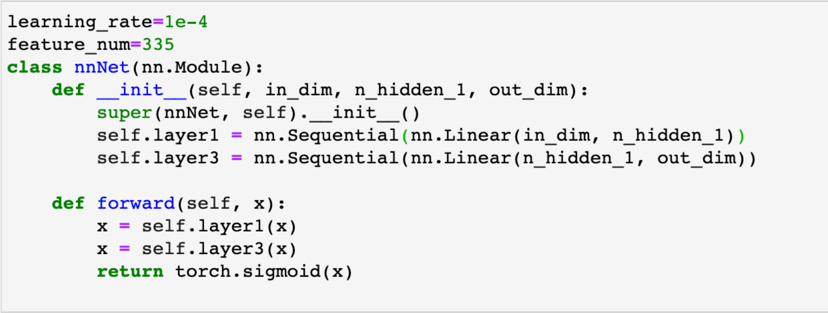


Figure 1: DOC2Vec Model

1.3 Learning Model

Here we use two Linear layers to train the model with hidden feature num=64. And then use sigmoid function to predict. And then we use BCEWithLogitsLoss to calcuate the loss of the model. The main part of the model is in 2



```
learning_rate=1e-4
feature_num=335
class nnNet(nn.Module):
    def __init__(self, in_dim, n_hidden_1, out_dim):
        super(nnNet, self).__init__()
        self.layer1 = nn.Sequential(nn.Linear(in_dim, n_hidden_1))
        self.layer3 = nn.Sequential(nn.Linear(n_hidden_1, out_dim))

    def forward(self, x):
        x = self.layer1(x)
        x = self.layer3(x)
        return torch.sigmoid(x)
```

Figure 2: Model Code

2 Results

2.1 Datasplit

Here we treat .ls data as test set. And we split the .train data given by the TA into train and valid data with frac=0.1. By this way, we can test whether our training should stop by calculating the roc-auc value of valid set.

2.2 Results

Here we try four situations: 1. use only base features in the paper(no). 2. use base features+doc2vec data for target region(local). 3. use base features+doc2vec data for the upstream and downstream data (global) 4. use base feauttres+doc2vec data for both target and upstream and downstream.(all)

Here is the results in table 1

name	all	no	local	global
C17ORF85_Baltz2012	0.864969610	0.870100244	0.862155655	0.878842055
C22ORF28_Baltz2012	0.851331333	0.854716	0.854828666	0.862462
CAPRIN1_Baltz2012	0.964473333	0.959435333	0.960881333	0.96396
CLIPSEQ_AGO2	0.855627333	0.853052	0.856152	0.852666
CLIPSEQ_ELAVL1	0.973195333	0.975478	0.973657333	0.974876666
CLIPSEQ_SFRS1	0.957561333	0.958650666	0.956665333	0.957765333
ALKBH5_Baltz2012	0.744214533	0.738089853	0.732417458	0.737147595

Table 1: Results

From the results, we can see that most of time, global performs better than no and local and all perform and local perform no better than no. So we can conclude that global one may not perform as well as only using base features. So we can conclude that the base information from the paper already contains all the information we need in this predicting work. And maybe some context information from global information will help the predicing work.

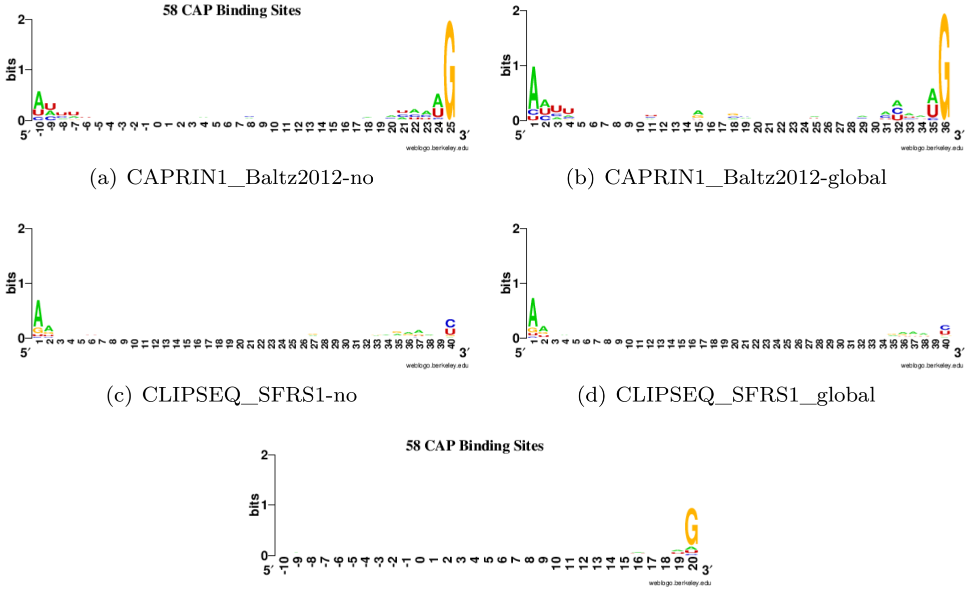
3 Transfer Learning

From the results, we can find that some tasks perform very well with high auc-roc value. But some task like ALKBH5_Baltz2012 don't perform very well. Therefore, I use transfer learning to train this task from the pretrained task CLIPSEQ_ELAVL1 using base features. And finally I get 0.771624830 result, increasing 4%.

But for the tasks like C17ORF85_Baltz2012 and C22ORF28_Baltz2012, the improvement is little or no. So we can only say that transfer perform well on some cases.

4 2019nCov prediction

Here we just use the average length of the targets in the training set. And select 250 symbols around the target. And then we train them find the sites which are nearly 1. Here are some pictures



(a) CAPRIN1_Baltz2012-no

(b) CAPRIN1_Baltz2012-global

(c) CLIPSEQ_SFRS1-no

(d) CLIPSEQ_SFRS1_global

(e) ALKBH5_Baltz2012-transfer

We can find that global results are more clear than the normal ones.

5 Conclusion

In this homework, I find that adding context information (doc2vec) doesn't improve the base features very well. And for some task with few data, we can use transfer learning to improve it. But because of the limitation of time, I haven't tried all the tasks.

And transfer learning can paly an important role in some tasks. This depends on the number of data and whether the data is noisy.

References

[1] Shuya Li, Fanghong Dong, Yuexin Wu, Sai Zhang, Chen Zhang, Xiao Liu, Tao Jiang, and Jianyang Zeng. A deep boosting based approach for capturing the sequence binding preferences of RNA-binding proteins from high-throughput CLIP-seq data. *Nucleic Acids Research*, 45(14):e129–e129, 05 2017.