COM 599100 Deep Learning Final Project Report – secondary structure (Q3,Q8) of a chain (Group 3)

First A. 104061136, Second B. **104061210** 104061132 107064518 105061129

ABSTRACT

Protein secondary structure prediction began in 1951 when Pauling and Corey predicted helical and sheet conformations for protein polypeptide backbone even before the first protein structure was determined. Sixty-five years later, powerful new methods breathe new life into this field. Deep learning has brought us to a new area in prediction recently and has also provided high precision results in the prediction of the secondary structure. In this paper, we treat the sequence of protein as a kind of language in NLP and use classic seq2seq model to predict Q3 structure as target language. We achieve 81% accuracy in test set, though the SOTA is 82.3% but our method prove that we don't need to preprocess the input protein sequence through complicate biological calculation like in the SOTA to get quite good performance.

I. INTRODUCTION

Deep NLP process has been proved being very powerful to solve sequence problem. In this project we treat predicting Q3 structure is like the task of machine translation, so we use the classical seq2seq model to solve the problem with the believing that the protein sequence is like one kind of language. Though we didn't beat the SOTA, however the SOTA requires very complicate biological computation and uses Deep Convolutional Neural Fields (DeepCNF) to do the task. Our approach is relatively quite simple and intuitive, in the future when we combine the knowledge and processing method of biology, we believe the approach will be very powerful.

II. MATERIAL AND METHOD

A. Dataset

The dataset we use contains protein sequence with different lengths and its correct secondary structures. The length of the sequence ranges from 3 to 5037. For data preprocessing, we first clean up the dataset by deleting those contain non-standard amino acids (*has_nonstd_aa = False*) and we get 386333 pieces of data. Then we randomly shuffled the whole dataset and choose **protein sequence** (column *seq*) **as source texts** and **Q3 structure** (*ssst3*) **as target texts**. For each text we choose 50000 pieces for training and testing.

2.2 Model

In this project, we use a classic seq2seq model(Fig 1)

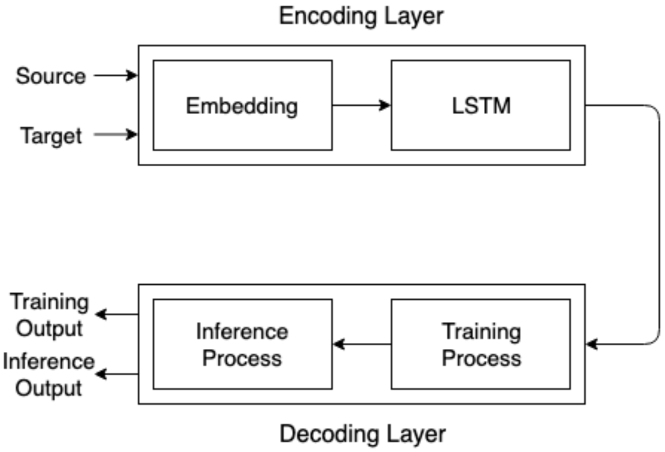


Fig. 1. Model Structure

I. Encoding Layer

We can separate encoding layer into two parts, embedding layer and RNN layers.

A. Embedding Layer

First, we create lookup tables (Fig. 2) for source text and target text. Both table are created with special token : <PAD>,

H.D. Jheng (104061210), Department of Electrical Engineering, National Tsing Hua University.

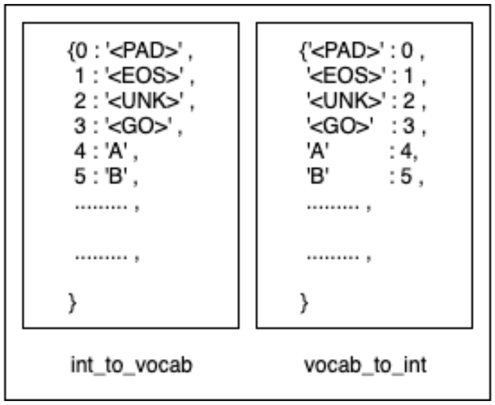
D.S. Chao (104061132), Department of Electrical Engineering, National Tsing Hua University.

S.H. Lin (104061136) , Department of Electrical Engineering, National Tsing Hua University.

Y.S. Yang (105061129), Department of Electrical Engineering, National Tsing Hua University.

B.H. Wu (107064518), Institute of Communication Engineering, National Tsing Hua University.

<GO>, <EOS>, <UNK> for padding, start, end of sentence, unknown signature respectively. The tables help transform the sequence into integers and can also map integer to chars. Later, using lookup tables to transform both source and target char to integer then embedding the integer sequence into 200 dimensional embedding latent space.



```
{0 : '<PAD>' ,
 1 : '<EOS>' ,
 2 : '<UNK>' ,
 3 : '<GO>' ,
 4 : 'A' ,
 5 : 'B' ,
 ......... ,

 ......... ,

}
```

int_to_vocab

```
{'<PAD>' : 0 ,
 '<EOS>' : 1 ,
 '<UNK>' : 2 ,
 '<GO>'  : 3 ,
 'A'     : 4,
 'B'     : 5 ,
 ......... ,

 ......... ,

}
```

vocab_to_int

Fig. 2. Lookup table

B. RNN Layers

Feeding the embedding result as input into LSTM with dropout (keep_prob = 0.5) for encoding the information of sequence. We can change the LSTM cell into any other RNN cell, such as GRU, etc.

II. Decoding Layer

Decoding model can be thought of two separate processes, training and inference. They share the same architecture and parameters, but they have different strategy to feed the shared model.

A. Training part (Fig. 3.)

The training part uses *dec_embed_input* as input, which is the output of *tf.nn.embedding_lookup(dec_embeddings, dec_input)*. This step manually creates embedding parameters for training phase to convert provided target data before the training is run.

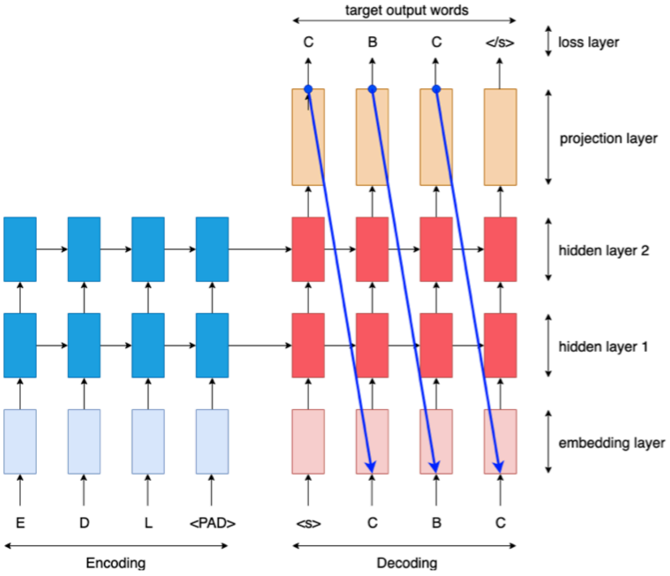


Fig. 3. Decoder – Training part

B. Inference part (Fig. 4.)

For the inference process, whenever the output of current time step is calculated via decoder, it will be embedded by the shared embedding parameters (the parameter we created above) and become the input of next time step. Because we need to share the parameter, so we need to define same variable scope and set reuse = True.

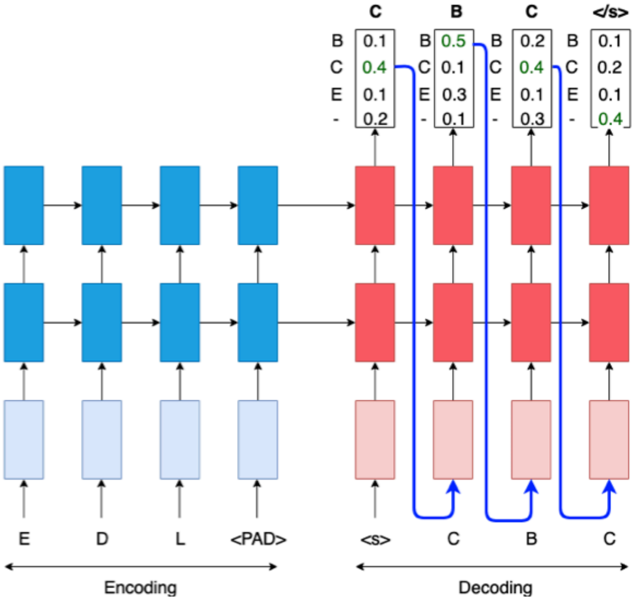


Fig. 4. Decoder – Inference part

III. Total Structure

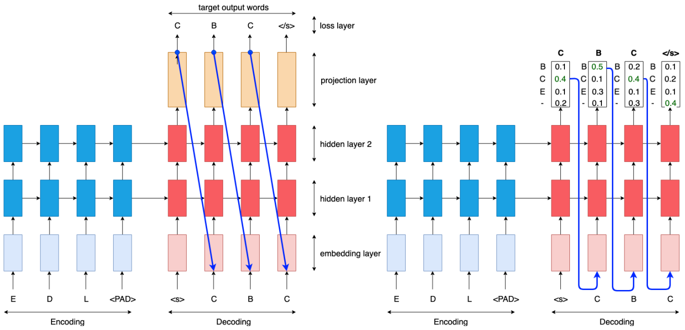


Fig. 5. Total Structure

III. Evaluation

A. Learning Curve

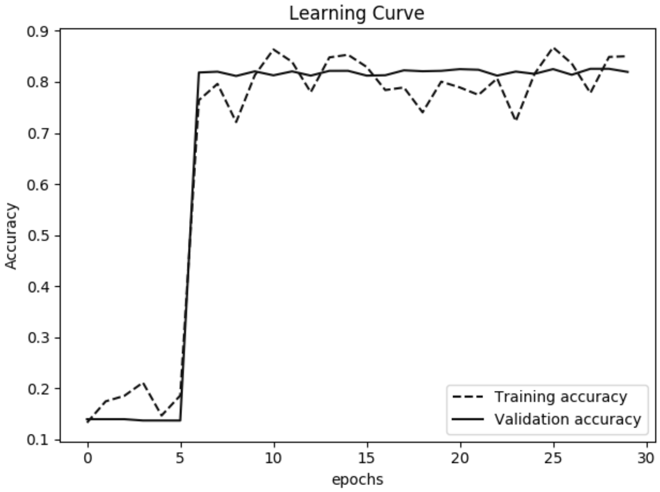
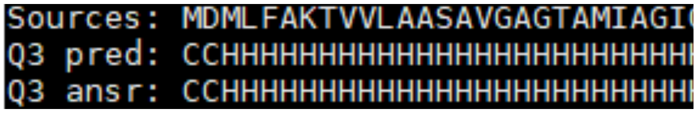


Fig. 6. Learning Curve

B. Result

We use accuracy to do the prediction. By calculating the difference between each element of prediction and correct answer. Below is the input sequence, target sequence and predicted result.



```
Sources: MDMLFAKTVVLAASAVGAGTAMIAGI
Q3 pred: CCHHHHHHHHHHHHHHHHHHHHHHHHH
Q3 ansr: CCHHHHHHHHHHHHHHHHHHHHHHHHH
```

Fig. 7. Result

IV. Discussion

We are surprised that by using a simple seq2seq model, we are able to predict the secondary structure without any biological computation and achieve 81% of accuracy. The accuracy is a little lower than the performance of SOTA. Maybe we can try bidirectional LSTM, unlike unidirectional LSTM, bidirectional LSTM preserves information from both past and future. And we can also try using pretrained embedding model such as ELMO, GLoVe, etc. to achieve better performance.

References

[1] Yuedong Yang, Jianzhao Gao, Jihua Wang, Rhys Heffernan, Jack Hanson, Kuldip Paliwal and Yaoqi Zhou. *In Sixty-five years of the long march in protein secondary structure prediction: the final stretch?*

[2] Sheng Wang, Jian Peng, Jianzhu Ma & Jinbo Xu. *In Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields*

[3] Ilya Sutskever, Oriol Vinyals, Quoc V. Le. *In Sequence to Sequence Learning with Neural Networks*

Work distribution

- ✧ Data preprocessing : 104061132
- ✧ Encoding : **104061210** 、105061129
- ✧ Decoding : 104061136 、107064518
- ✧ Merged & Train : 105061129