Homework 1

Chinese Text Segmentation Using BI-LSTM

Student Name: Zubair Baqai

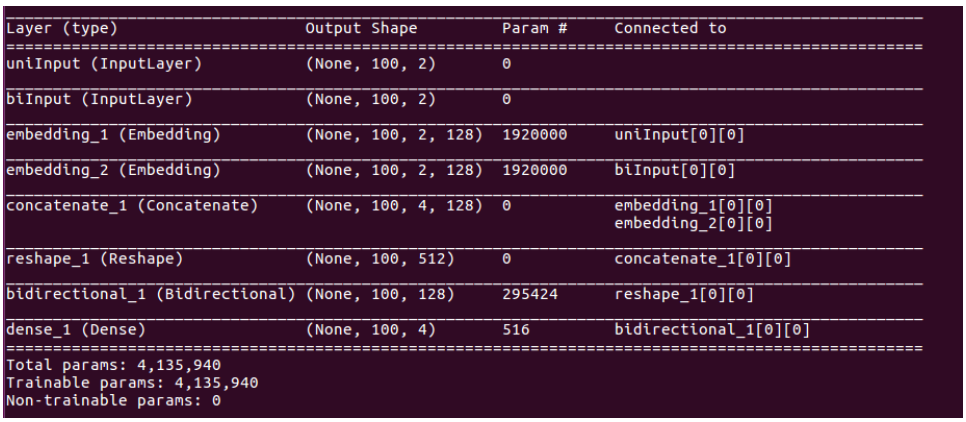
Matricola # : 1849940

Project report

This project aims to solve the Chinese language Segmentation problem , Many Researched and Breakthroughs are being made in Recent years .
We Have Decided to use Neural network to solve this problem , Main Layers in this neural network are following

MODEL

- We Use two **Embedding Layers** , One for Unigrams and one for bigrams . Both layers Take an input of shape [100,2] .and result Embedding values.
- After getting the Emeddings, we use **Concatenation Layer** , to merge the outputs from the embedding layer .
- After we have merged the Embeddings , we must Reshape the Inputs , Because BiLSTM dosnt take 3d inputs , so we use **Reshaping layers** to do the job
- After the inputs are Reshaped we Pass it into the **Stacked LSTM layer,** which move in opposite direction , as suggested in Research paper
- Finally we pass the outputs from LSTM layer into TimeDistributed Dense layer which has Softmax Activation Function.



Layer (type)	Output Shape	Param #	Connected to
uniInput (InputLayer)	(None, 100, 2)	0	
biInput (InputLayer)	(None, 100, 2)	0	
embedding_1 (Embedding)	(None, 100, 2, 128)	1920000	uniInput[0][0]
embedding_2 (Embedding)	(None, 100, 2, 128)	1920000	biInput[0][0]
concatenate_1 (Concatenate)	(None, 100, 4, 128)	0	embedding_1[0][0] embedding_2[0][0]
reshape_1 (Reshape)	(None, 100, 512)	0	concatenate_1[0][0]
bidirectional_1 (Bidirectional)	(None, 100, 128)	295424	reshape_1[0][0]
dense_1 (Dense)	(None, 100, 4)	516	bidirectional_1[0][0]

Total params: 4,135,940
Trainable params: 4,135,940
Non-trainable params: 0

Hyper Parameters

After Testing the Model with Different combination of HyperParameters , I found the following paramters to be the Best resulting once

- Emedding Layer of Both Unigram and Bigram are set to 128
- For LSTM , we use 64 cells , and We Set Drop and Recurrent Dropout to be 0.2
- We Set the Learning Rate to be 0.04 ,Decay rate to be 1e-5, Momentum 0.95

Data Sets

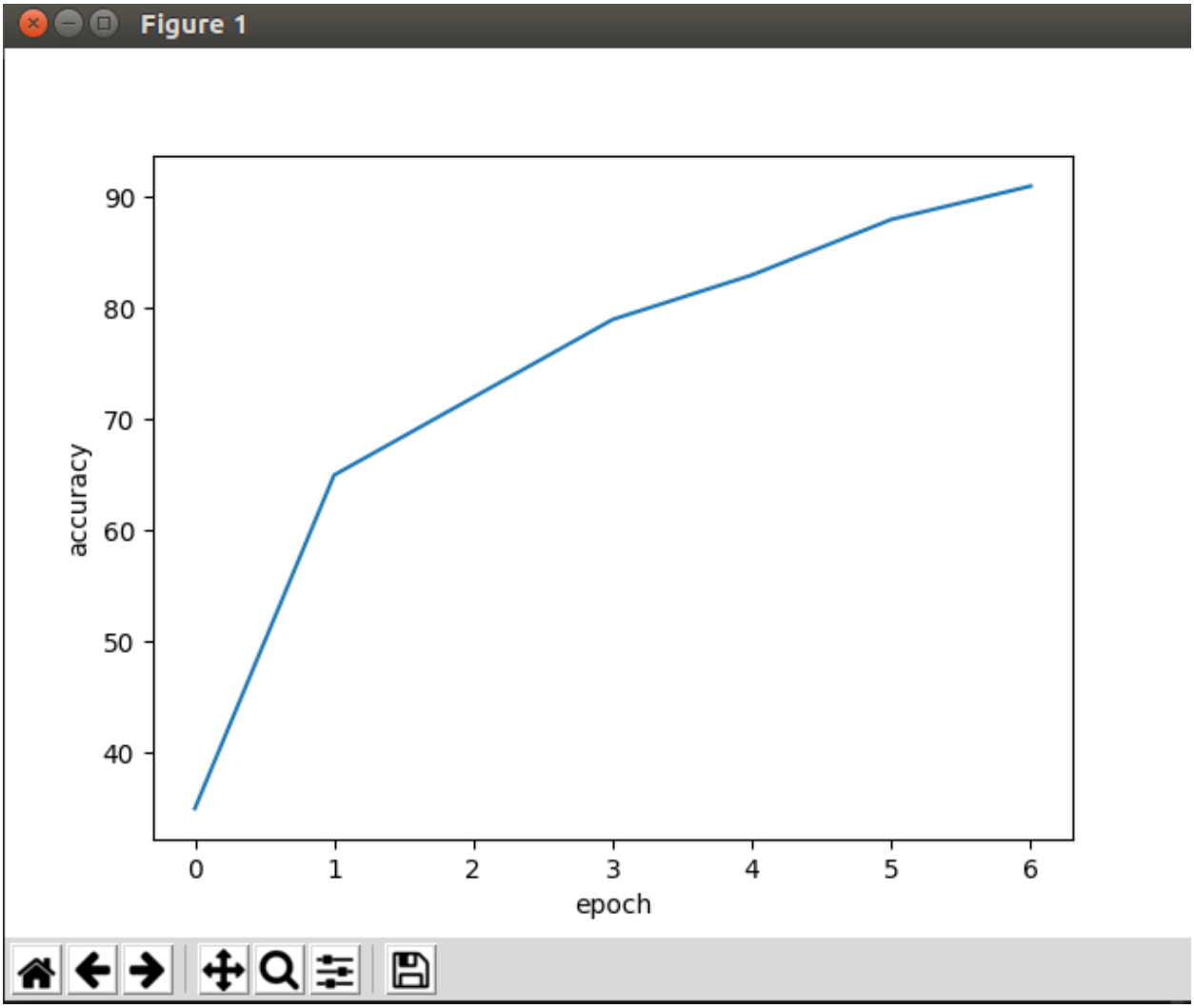
Firstly, We Made sure all the Datasets are being Converted to Simplified Chinese using the Hanzi-Convert . Then using these converted text, We have Concatenated all the Datasets , removed the Spaces and NewLines for training. Furthermore we have chosen to feed the Model with an Array of 100 Characters each for Unigrams and Bigrams. All our inputs are being converted to Integers as Neural network understand Integers rather than the characters or string . We also have Generated the labels using the Preprocessing , which are then converted to One-hot Encoding , so Softmax can Classify the inputs amongst the four label(B,I,E,S) .

Improvements

My first attempt was made to send sentences and add padding or truncate to it if the sentence were not as of the Required Length , but after testing it , I found the accuracy was not enough and a lot of overfitting could be seen. So as a solution I designed the algorithm that it treats complete Dataset as single Sentence , and break it into chunk of 100 . And if last row had les then 100 we

just add padding to last row . Which significantly Improved the overall performance and accuracy .

Plots



This is the precision I got after running prediction on a complete untrained data found in gold testing .

```
baqai@baqai-Alienware-15-R3:~/Desktop/code/public_homework_1/Final code$ python3
 score.py "/home/baqai/Desktop/code/public_homework_1/icwb2-data/training/as_Lab
el1.utf8" "/home/baqai/Desktop/code/public_homework_1/icwb2-data/training/as_out
put1.utf8"
Precision:      0.9254171462048454
baqai@baqai-Alienware-15-R3:~/Desktop/code/public_homework_1/Final code$
```

References

1. Ma, Ji & Ganchev, Kuzman & Weiss, David. (2018). State-of-the-art Chinese Word Segmentation with Bi-LSTMs.