iNetworks Machine Learning Internship

Job Classifier

Using Keras and LSTM

Abdulrahman AbouOuf

1-4-2020

CONTENTS

TABLE OF FIGURES

1 DATA CLEANING

Only used first two columns Job Title and Job Function

1.1 CLEANING JOB FUNCTIONS

Only Cleaning job functions removing / - and spaces so it is easier for the classifier.

```
def cleanjobfun(jobfuns):
  jobfun_cleaned = []
  for jobfun in jobfuns:
    for word in jobfun:
      word = word.replace("/" , "")
      word = word.replace("-" , "")
      word = word.replace(" " , "")
      if(word !='nan'):
        jobfun_cleaned.append(word)
  return list(dict.fromkeys(jobfun_cleaned))
```

Figure 1-1 Cleaning Function

x	y
X_1	$[t_2, t_5]$
X_2	$[t_1, t_3, t_2]$
X_3	$[t_4]$

x	t_1	t_2	t_3	t_4	t_5
X_1	0	1	0	0	1
X_2	1	1	1	0	0
X_3	0	0	0	1	0

1.2 MULTILABEL CLASSIFIER

As the job title produces multiple job functions. I used Multi-Label Classifier

```
multilabel_binarizer = MultiLabelBinarizer()
multilabel_binarizer.fit_transform(job_dataframe_cleaned['jobFunction'])
y = multilabel_binarizer.transform(job_dataframe_cleaned['jobFunction'])
```

Figure 1-2 Multiple Classifier in y

2 MODEL

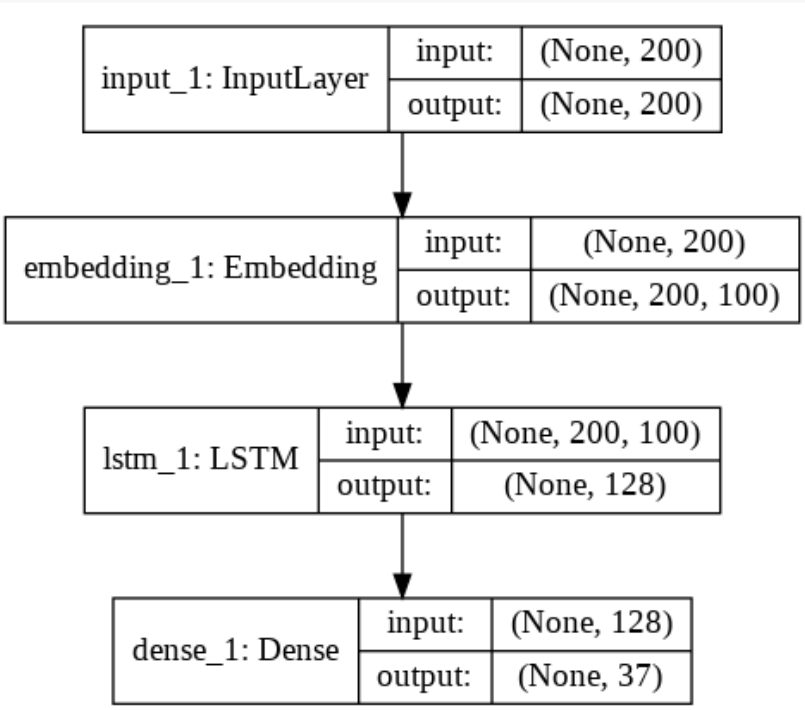

Figure 2-1 Model Pipeline

2.1 LTSM

During the training of RNN, as the information goes in loop again and again which results in very large updates to neural network model weights. This is due to the accumulation of error gradients during an update and hence, results in an unstable network. At an extreme, the values of weights can become so large as to overflow and result in NaN values.The explosion occurs through exponential growth by repeatedly multiplying gradients through the network layers that have values larger than 1 or vanishing occurs if the values are less than 1.

The above drawback of RNN pushed the scientists to develop and invent a new variant of the RNN model, called Long Short Term Memory. LSTM can solve this problem, because it uses gates to control the memorizing process.

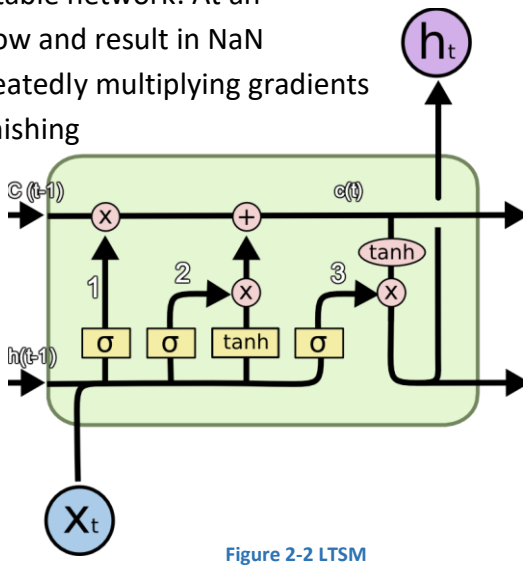

Figure 2-2 LTSM

2.2 GLOVE

To embed words of similar meaning to same vectors, more in the notebook.

3 WHY THIS MODEL?

As discussed in the previous section, this is faster and easier to implement with text. For move visit https://stackabuse.com/python-for-nlp-multi-label-text-classification-with-keras/#disqus_thread

4 BETTER PERFORMANCE?!

Can use another threshold to train and classify with more word embedding. Or using simple ML model.

5 EVALUATION

```
611/611 [==============================] - 3s 5ms/step
Test Score/ F1 Score: 0.17891554729540493
Test Accuracy: 0.9444419911372096
```

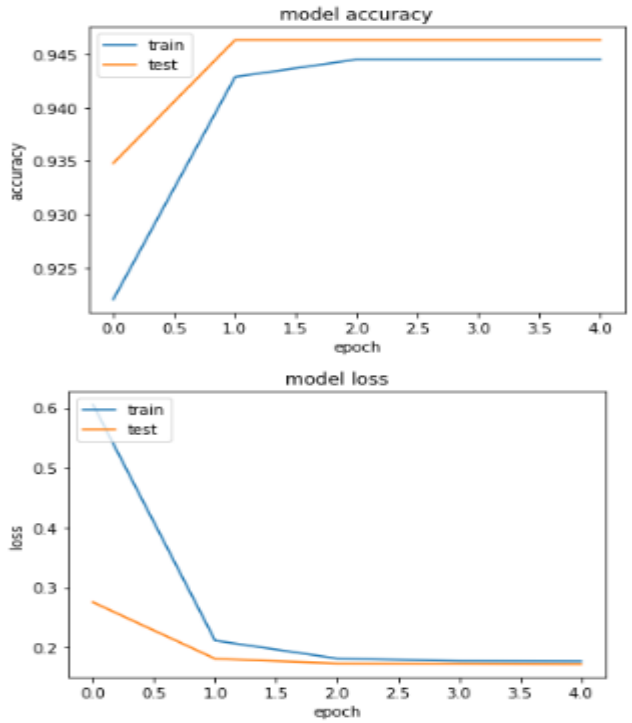

Figure 5-1 Model History

6 LIMITATIONS

Not enough Description in job title. And using Arabic language might fail in word embedding.

7 REFERENCES

- https://nlp.stanford.edu/projects/glove/
- https://www.analyticsvidhya.com/blog/2019/04/predicting-movie-genres-nlp-multi-label-classification/
- https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MultiLabelBinarizer.html
- https://towardsdatascience.com/understanding-lstm-and-its-quick-implementation-in-keras-for-sentiment-analysis-af410fd85b47
- https://stackabuse.com/python-for-nlp-multi-label-text-classification-with-keras/