Journal of Electronic Research and Application

Research Article

A Sentimental Analysis System for Film Review based on Deep Learning

Keyao Wu*

The High School Affiliated to Renmin University of China

Abstract: The paper will be introduced as sentimental analysis system of film criticism based on deep learning. Which contains four main processing sections. Compared with other systems, our sentimental analysis system based on deep learning has plenty of advantages, including simple structure, high accuracy, and rapid encoding speed.

Keywords: Deep learning, Data processing, Convolutional Neural Networks, Sentimental analysis system

Publication date: September, 2019

Publication online: 30 September, 2019

Corresponding author: Keyao Wu, liaoquanneng@ivygate.cn

1 Introduction

With the explosive growth of this type of comment, the demand of the technology of the sentimental analysis, a section of the natural language processing (NLP), are gradually increasing, as it can be employed to analyze and judge the emotional types of text description, so that the machine can be prone to comprehend the emotions and views expressed in the text. Nonetheless, due to both the complexity and diversity of the human languages, the applications of the sentimental analysis are considered as a challenging task.

Previous researches show that the basic machine learning techniques can accomplish some natural language processing tasks effectively, such as document subject classification. However, the same techniques cannot be applied to the field of emotional classification, since it requires more efforts to overcome the challenges emotional analysis faced and to deal with the diversity involved in emotional analysis.

Accordingly, we decide to adopt convolutional neural network (CNN) and fully connected neural network (FC) in our program of emotional analysis of film reviews. With its special structure of local weight sharing, convolutional neural network has unique advantages in speech recognition and image processing. Its layout is closer to the actual biological neural network. We take advantage of 4 convolution layers, 2 max pooling layers, and 2 full connection layers, which not only makes the overall framework structure of the project relatively simple and speeds up, but achieves the content that the previous design wants to optimize as well.

2 Major steps of sentimental analysis system

A sentimental analysis system for film review based on deep learning includes six major steps:

1) Firstly, the original film review data would be preprocessed, which includes eliminating html tags; deleting non-character information; and utilizing the nltk.stopword in python to cast off the stop words.

2) Then our system would transform the preprocessed data to the form of Bag-of-words Model, which serves to transform natural language information to arrays conducted with numbers.

3) The system would then transfer the sentiment in the data into the form of one-hot-encoding. While the system is executing the Bag-of-words Model transformation, it would label the Bag-of-words from each review with their corresponding sentiment.

4) The data after such process, along with the sentiments in the form of one-hot-encoding, would be imported as input samples and labels into the deep learning network in our system.

5) And then we use primarily three methods for the

optimization of the learning structure.

6) Last but not least, we split the film review data to the training set and testing set in at a ratio of p:q. The system uses the training set as input data to train and modify our emotional analysis model and uses the testing set to calculate the accuracy of the model.

3 Processing and verification

3.1 Data processing & one hot encoding

As our data is given in the form of a tsv file, we must import a library to open such a file. The best choice and our choice are no doubt the "pandas". We applied append method within the python interpreter to put the preprocessed data into an empty list "a" as we defined before. The following task would be turning the labels, in this case, the "sentiment" column of the data, into one-hot encoding labels.

3.2 Bag of words

After the success of data cleaning, we conducted data processing and feature construction. We use the bag-of-words to construct text features and it was originally used in the field of information retrieval. We took 996 reviews and broke them down into individual words. Then we culled the top 5,000 words. Make these five thousand words as a dictionary. Using any order listed in our dictionary, we can convert reviews to binary vectors. All sorts of traditional document-like words are discarded, and we can use this generic method to extract features from any document in our corpus, which can then be used for modeling.

3.3 Convolutional neural network

We used neural network, which comprises of four convolutional layers, two pooling layers, and two fully connected layers, to train our models. Except for neural network, random forest is also a commonly used machine-learning algorithm; we will discuss their differences and explain why our final option is neural network. First of all, the random forest algorithm is rather independent and conventional, which holds for every one of its decision tree whereas the neural network is closely bonded with all of its neurons, one cannot work without each other. Secondly, random forest algorithm can only process data provided in chart form which would have cost us a lot of inconveniences if we had used this.

3.4 Optimize

The last thing we need to do is optimize this program. The whole has two aspects. The first is the optimization of the whole code, and the second is the optimization method for the design architecture. When we actually did the convolution, we found that each convolution layer was coded with a different name, and the contents were the same. By comparing a gif of parameter updating methods, Adam is the fastest method with the least error.

4 Conclusion

1. Since we choose the top 5000 high frequency words, only 2 MaxPooling layers are needed, which led to our structure relatively simple with higher processing speed.

2. We use a structure with high accuracy, convolutional neural network (CNN), to build up our system. Consequently, with appropriate parameters, our system can maintain a relatively high accuracy (about 84%).