International Core Journal of Engineering Volume 8 Issue 2, 2022
ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0047

News Text Data Classification Method based on Deep Learning Algorithm

Weikun Qiang[1, *], Ziyuan Jiang[2, a], Xinkai Zhou[3, b], Yanyu Zhu[4, c], Keyu Zhu[5, d]

[1] Soochow University, Suzhou, Jiangsu, China

[2] University of Reading, Reading, UK

[3] Zhejiang Zhuji Middle School, Zhuji, Zhejiang, China

[4] Claremont School, East Sussex, UK

[5] Zhejiang Huawei Foreign Language School, Shaoxing, Zhejiang, China

[a]bj803118@student.reading.ac.uk, [b]1994268836@qq.com, [c]uyanyu492@163.com, [d]1416482149@qq.com

[*]Corresponding author: 1320272580@qq.com

These authors contributed equally to this work

Abstract

News and information app is an important channel for news content distribution. In order to meet the different needs of different groups of people, news classification is crucial. Using automatic classification can reduce a lot of manual operation, which is a core work for the formation of information flow recommendation.This paper proposes CNN(convolutional neural network) to design news classification analysis) to solve this problem.In the daily exposure of people's news content including sports, finance, real estate, home, education, technology, technology, fashion, current politics, games, entertainment, so we decided to experiment with the data of these ten categories of news content.In the method we designed this time, they are mainly divided into four steps: loading data, processing text, text vectorization, and forming text matrix, so that the data are divided into training samples for modeling, evaluating samples for adjusting parameters, and testing samples for evaluating the model.After processing over 650,000 experimental data using python3, we showed that CNN achieved 96% accuracy and RNN achieved 94% accuracy.

Keywords

Deep Learning Algorithm; Automatic Classification; CNN; Python3.

1. Introduction

At the end of the last century, with the rise of the Internet, the rapid development of network technology, which set off a wave of network, and more and more people can have access to the network and use the network.The number of global online surged, and our lives have been surrounded by digital information.In today's rapid development, people's demand for information is more and more intense, and now in a competitive and fast-paced life, fragmented reading has become the most converging way in the tide of The Times.In 2007, smartphones began to truly enter the market and became popularized to a certain extent. News app s also quickly entered the lives of many people with smartphones.For users of different ages, professions, and genders, most news app classify their huge amount of information into many categories and recommend users news they may be interested

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0047

in.In this whole link, the processing of news and information is very important.If such a large amount of data processing is all handed over to manual completion, for both human and financial consumption is immeasurable.Therefore, at this time, information technology needs to complete the task of automatic classification, so as to successfully complete the information flow recommendation.Our study mainly designed news classification analysis using CNN(convolutional neural network) and RNN(recurrent neural network).Chinese text classification was performed using convolutional, recurrent neural networks based on the TensorFlow deep learning framework.We first of all the ordinary people will contact the news information roughly divided into sports, finance, property, home, home, education, technology, fashion, politics, games, entertainment and other 10 categories.And experiment with the data of these ten categories of news content.In the experiments we mainly used python3 as the main tool, with TensorFlow 1.3 above and numpy,scikit-learn,scipy as the main experimental environment.Design method steps include loading data, processing text, text vectorization, and forming a text matrix.We divided the Chinese text dataset used for this training into 10 classifications, each containing 65,000 data bars.We divided all the data into training, validation and test sets, training samples used for modeling, evaluating samples for adjusting parameters, test samples for evaluating the model, from the original data forming a subset and integration execution to produce three data files:

cnews.train.txt: training set (50,000 bars).

cnews.val.txt: Validation Set (5,000 bars).

cnews.test.txt: test set (10,000 bars).

The resulting data were then preprocessed to produce the CNN(convolutional neural network) and the RNN(recurrent neural network) model.Finally training finally trained through python.

2. Design

The CNN convolutional neural network algorithm judges different objects by extracting the local features from the original data and then passing up and combining to obtain the overall features.CNN has the following advantages: First, processing images usually requires large computation and excessive time costs, and CNN simplifies the computational complexity by reducing the raw data dimension.Secondly, the original data is digitized and it is difficult to fully preserve the characteristics, which will lead to the classification accuracy of the model. CNN extracts the local characteristics of the image at the same time when processing the data, so that the accuracy of the model classification is guaranteed.The advantage of the image domain CNN also exists in processing text semantic classification, compared to the traditional model CNN algorithm that can sense the semantic association of two words more located away in the text through the underlying local features.

First, the data is first converted into a vector that the computer can process via the embedding layer (embedding).The processed data is then passed via the CNN convolutional neural network (including convolutional, pooling, fully connected layers).Convolutional layers in the CNN model are mainly scanned by convolution checking requiring the text of the extracted features to obtain the eigenvalues for these regions.Since the data features processed by the convolutional layer are still very large, it also needs to be processed by the pooling layer to reduce the dimensionality, which can reduce the complexity of subsequent operations and avoid the occurrence of overfitting.Data features processed by the embedding layer, convolutional layer and pooling layer enter the full connected layer, and initial output results are finally obtained through the input layer, hidden layer and output layer.Finally, the output of the neurons was mapped to the interval of 0 to 1 by the softmax function to facilitate the classification of the data.For example, in this application, the channel data of the raw text results through successive convolution of the convolutional kernel, yielding the output of the convolutional layer in the order that the convolutional kernel moves.In processing the raw data, the data needs to undergo multiple layers of convolution and pooling to reach the desired structure.The convolutional data reduce their dimensionality through pooling layers, simplify the complexity of subsequent operations, and also increase the stability of features within the model.The steps of the pooling layer

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0047

is similar to the convolutional layer, and the filter of the pooling layer pools the output data of the convolutional layer.Finally, the classification result was obtained with multiple training of the fully connected layers.

Among these, the parameters of the CNN convolutional neural networks are as follows, and embedding_dim this parameter determines the vector dimension of the text in the embedding layer, assigned 64 in this model. Both sequence length and number of categories are parameters set upon request, seq_length = 600 num_classes = 10, respectively.The two parameters of num_filters = 128 kernel_size = 5 define the number kernel size of convolutional kernels, respectively.hidden_dim = 128 This parameter determines the dimensionality of neurons in the fully connected layer.However, the dropout parameter improves the stability and robustness of the model, in which it is chosen as 0.5. Furthermore, the learning rate and training size in the model selected learning_rate = 1 e-3 and batch_size = 64 by adjusting the parameters and the properties of the reference CNN. Through subsequent training, the above two parameters are more reasonable, with no excessive gradient or gradient disappearance.

Also, we also used the RNN model.Unlike the CNN model, RNN is better suited to handling problems in the field of natural language processing.For example, when dealing with natural language, in a sentence, the order of words appearing has a great impact on the prediction of the entire sentence of semantic language.Second, the data entered by each node of the RNN contains all the data from the previous layer.Therefore, the RNN model would be more suited when processing information about a sequence because the RNN model considers the effect of the order of word appearance.RNN is a class of memory-capable neural network models, which can only have short memory power for each node due to gradient explosion or gradient disappearance.In the neural networks of RNN, neurons between layers also establish weight-related connections, different from traditional basal neural networks that only make connections between layers.Moreover, in the structure of RNN, there are also weights between neurons in the hidden layers, meaning that the previous hidden layers influence layers later as the neural network advances.In the parameter settings of the RNN recurrent neural network, the parameters regarding the embedding and softmax layers are the same as in the CNN.After considering the stability and robustness of the model, the dropout parameter was chosen as 0.8.

In the RNN model that handles text semantics, the raw text data first reduces dimensions in the embedding layer.What is different from CNN in subsequent neural networks is that the individual neurons in the RNN neural network associate not only with the input and output, but also with themselves.The state of their own neurons at the last moment acts on the neuronal state at the next moment, which means in text processing that the anterior and posterior words of a sentence can be associated in the RNN neural network.This process, also called the positive propagation of RNN, is associated in most cases of the anterior and posterior sentences in natural language, and this positive propagation characteristic will undoubtedly improve the accuracy of RNN network classification.

3. Experimental Results

This training used 10 of them, respectively for sports, finance, real estate, home furnishing, education, science and technology, fashion, current politics, games, entertainment, each classification has 65,000 pieces of data.

This experimental data is one subset of data, datasets:

Training set: 50,000 * 10 (Estimated model).

Validation set: 5000 * 10 (parameters determining network structure or controlling model complexity).

Test set: 10000 * 10 (test the performance of the best model finally selected).

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0047

For the procedure of generating a subset from the original dataset, see two scripts under helper.Among them, the copy_data.sh was used to copy the 6,500 files per classification, cnews_group.py is used to integrate multiple files into one file.After executing the file, data files are three:

cnews.train.txt: training set (50000).

cnesw.val.txt: validation set (5000).

cnews.test.txt: test set (10000).

Chinese text classification was performed using convolutional, recurrent neural networks based on the TensorFlow deep learning framework.(1) Environment:

a) Python 3.

b) TensorFlow 1.3 Up.

c) numpy.

d) scikit-learn.

e) scipy.

(2) Preprocessing of the datasets.

data/cnews_loader.The py is a preprocessing file for the data.

read_file (): Read the file data;

build_vocab (): Build a vocabulary, using character-level representations, this function stores the vocabulary to avoid.

Each treatment was repeated;

read_vocab (): Read the vocabulary stored last step to {word: id};

read_category (): Fixed the classification directory to {Category: id} representation;

to_words (): Revert a data represented by id to text;

process_file (): Converts datasets from text to a fixed-length id sequence representation;

batch_iter (): Prepare the shuffle batch of data for the training of the neural network.

After data preprocessing, the data format as follows:

Data	Shape	Data	Shape
x_train	[50000, 600]	y_train	[50000, 10]
x_val	[5000, 600]	y_val	[5000, 10]
x_test	[10000, 600]	y_test	[10000, 10]

Figure 1. The data format

This test of CNN convolutional neural network and RNN recurrent neural network, we run python.run_cnn.py train can start training (ps: if trained previously, remove tensorboard /textcnn to avoid repeating multiple TensorBoard training results).

The best effect of CNN after 3 iterations on the validation set was 94.12% with the accuracy and error shown in Fig,

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0047

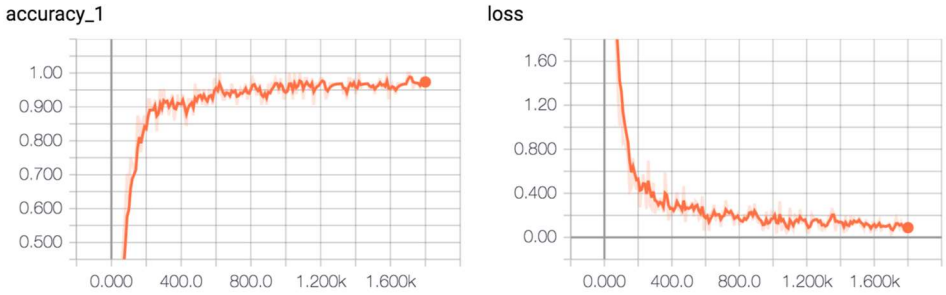

Figure 2. The accuracy and error of CNN after 3 iterations

The best effect of RNN after eight rounds of iterations on the validation set was 91.42%, much slower than CNN, with the accuracy and error shown in Fig,

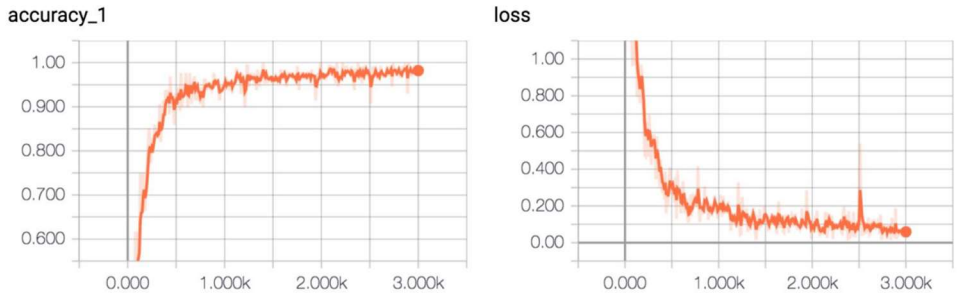

Figure 3. The accuracy and error of RNN after eight rounds of iterations

Then you run the python run_cnn.The py test is tested on the test set:

Configuring CNN model...

Load i ng test data

Test Loss 014, Test Acc 96.04%

Test i ng

Precision, Recall and F1Score...

Table 1. The test data of CNN model

	precision	recall	f1-score	support
Sports	0.99	0.99	0.99	1000
Finance and economics	0.96	0.99	0.97	1000
Property	1.00	1.00	1.00	1000
Home	0.95	0.91	0.93	1000
Education	0.95	0.89	0.92	1000
Technology	0.94	0.97	0.95	1000
Fashion	0.95	0.97	0.96	1000
Current politics	0.94	0.94	0.94	1000
Game	0.97	0.96	0.97	1000
Entertainment	0.95	0.98	0.97	1000
avg/total	0.96	0.96	0.96	10000

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0047

According to the data, the accuracy of CNN reached 96.04% on the test set, and various precision,recall and f1-score exceeded 0.9, where the performance in the political category was poor but also reached 0.94, and 1.00. We can also see from the confusion matrix that the classification effect is very excellent.

Run the python run_rnn.The py test was tested on the test set.

Test Loss: 0.21 Test Acc: 94.22%.

Testing.

Precision, Recall and F1-core.

Table 2. The test data of RNN model

	precision	recall	f1-score	support
Sports	0.99	0.99	0.99	1000
Finance and economics	0.91	0.99	0.95	1000
Property	1.00	1.00	1.00	1000
Home	0.97	0.73	0.83	1000
Education	0.91	0.92	0.91	1000
Technology	0.93	0.96	0.94	1000
Fashion	0.89	0.97	0.93	1000
Current politics	0.93	0.93	0.93	1000
Game	0.95	0.97	0.96	1000
Entertainment	0.97	0.96	0.97	1000
avg/total	0.94	0.94	0.94	10000

Comparing the results of CNN, RNN achieved 94.22% accuracy on the test set, also with high accuracy, and various precision,recall and f1-score exceeded 0.9, but individual data are not very ideal (home recall and f1-score, fashionable precision), with good results of 1.00 on property. Taken together, RNN can also achieve better results through parameter tuning.

4. Conclusion

In this paper, a convolutional neural network is used to build a classification model of news content, and the experimental results show that this method is effective. According the data, CNN achieved 96.04% accuracy on the test set, and from the confusion matrix, the classification is also excellent. Comparing the results of CNN, RNN achieved 94.22% accuracy on the test set, also with high accuracy. The results of this paper can be used for the recommendation classification of news content APP and short video APP like information streaming recommendations.

References

[1] Tomoya Matsumoto, Wataru Sunayama, Yuji Hatanaka, Kazunori Ogohara. Data Analysis Support by Combining Data Mining and Text Mining, 2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI).

[2] Ning Zhong, Yuefeng Li, Sheng-Tang Wu. Effective Pattern Discovery for Text Mining, IEEE Transactions on Knowledge and Data Engineering, 2012, vol.24(1).

[3] Deepak Agnihotri, Kesari Verma, Priyanka Tripathi. Pattern and Cluster Mining on Text Data. 2014 Fourth International Conference on Communication Systems and Network Technologies.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0047

[4] Bin Zhou, Yan Jia, Chunyang Liu, Xu Zhang. A Distributed Text Mining System for Online Web Textual Data Analysis, 2010 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery.

[5] Weitao Weng, Yongbin Liu, Sibo Wang, Kai Lei. A multiclass classification model for stock news based on structured data, 2016 Sixth International Conference on Information Science and Technology (ICIST).

[6] Ibrahim R. Hallac, Betul Ay, Galip Aydin. Experiments on Fine Tuning Deep Learning Models With News Data For Tweet Classification, 2018 International Conference on Artificial Intelligence and Data Processing (IDAP).

[7] Fan Zhang, Wang Gao, Yuan Fang. News Title Classification Based on Sentence-LDA Model and Word Embedding, 2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI).

[8] Zhenzhong Li, Wenqian Shang, Menghan Yan. News text classification model based on topic model, 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS).

[9] Shenhao Zhang, Yihui Wang, Chengxiang Tan, Research on Text Classification for Identifying Fake News, 2018 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC).

[10] Metin Mert Akçay, et al. Sport News Classification with Convolutional Neural Network and Long-Short Term Memory, 2019 27th Signal Processing and Communications Applications Conference (SIU).

[11] Jiangnan Qi, Yuan Rao, et al. Semantic Enhancement and Multi-level Label Embedding for Chinese News Headline Classification, 2019 14th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP).

[12] Benjamin D. Horne, William Dron, Sibel Adali. Models for Predicting Community-Specific Interest in News Articles, 2018 IEEE Military Communications Conference (MILCOM).