The Research Progress of the Deep Hybrid Model in the Field of Text Classification

Chunfeng Yang¹, Qiang Wu¹*, Jiajia Lu¹ and Huiyu Chen¹

¹ School of computer science and Engineering, Northwest Normal University, Lanzhou, GanSu Province, 730000, China

*Corresponding author’s e-mail: 2019221794@nwnu.edu.cn

Abstract. The traditional deep convolution neural network model cannot extract the context information effectively in dealing with complex long text data sets, and it is difficult to obtain the deep semantic information of the text. The deep hybrid neural network can optimize and improve the local feature extraction ability of CNN model while preserving the ability of local feature extraction. It has achieved good performance on complex data sets, so it has been paid more and more attention by researchers. Firstly, this article sorts out the current mainstream text classification data sets. Secondly, the model of the hybrid neural network based on the convolutional neural network construction is as follows: the improvement of the CNN model; technology fusion based on CNN model and the CNN-based model mixing three categories carry out analysis and sorting. Finally summarize the current problems in the text classification field, and look forward to future development and research.

1. Introduction

As a method of information management, the main purpose of text classification is to classify text according to certain rules. This “rule” can be determined by people or automatically summarized by algorithms from labeled data. Text classification plays an important role in emotional recognition, [1] social opinion classification [2], intelligent question and answer [3], natural language inference [4] and so on. In the early days, Bengio and others used word vectors to classify text and proposed neural network language model NNLM [5]. In 2014, google proposed a more concise language model framework Word2vec [6][7] based on bengio to generate word vectors, text classification based on neural network is becoming more and more popular. When the traditional machine learning method is applied to text classification, it mainly extracts the features of the input text, and then uses the classifier for text classification. More representative feature extraction methods are: One-hot, Bag of words, TF-IDF [8] etc. On this basis, the use of simple Bayesian classifier, support vector machine and other machine learning [9] methods can achieve good classification results. With the development of deep neural networks, researchers began to work on the application of deep learning to text classification, and until 2014, Kim improved convolutional neural network, which is often mentioned in the field of machine vision, and proposed text classification model based on it (TextCNN) [10]. Compared with the traditional image CNN network [11], TextCNN keeps the original network structure and simplifies the convolution layer. The network structure of TextCNN is simple, so the number of parameters is small, the amount of calculation is small, and the training speed is fast, so it's hard to assess the importance of each feature. In view of these shortcomings, later researchers have carried out a great degree of transformation on TextCNN, put forward many complex hybrid models based on convolutional neural network, and applied them to practice text classification datasets for different application areas.
2. Text classification data sets in different application fields
The application of deep learning in the field of text classification is more and more popular, at the same time, it also promotes the development of text classification data sets. The performance of the same deep learning model in different data sets must be different, so how to select the appropriate neural network model for different text classification data sets is also the focus of future research, the data set is very complex, including many excellent public data sets, including some data sets constructed by individuals or units. Because of the differences between data sets, it is necessary to have a basic understanding of the mainstream data sets. Table 1 shows the mainstream data sets that are widely used in the field of text classification in recent years.

| Applications          | data set       | Brief introduction                                                                                                                                                                                                 |
|-----------------------|----------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Emotional analysis    | Yelp           | Contains two types of data: one is a label that detects fine-grainedness, called Yelp-5. Another predictor of negative and positive emotions is called Yelp-2.                                                        |
|                       | IMDb           | Developed for the second classification emotional classification task of movie reviews. by and so on the composition of positive and negative comments. The training and test sets are evenly distributed, with 25,000 comments each. |
|                       | Movie Review   | A collection of movie reviews whose purpose is to detect emotions associated with a particular review and determine whether it is negative or positive. Includes 10,662 sentences with negative and positive samples, typically using 10x cross-validation and random splitting to test the dataset |
|                       | SST            | It contains two types of data: one with a fine-grained label (five categories) and the other with a two-category label called SST-1 and SST-2, respectively.                                                                   |
| News classification   | AG-News        | A collection of news articles collected by academic news search engine. Come to My Head from more than 2,000 news sources.                                                                                                  |
|                       | 20 newsgroups  | A collection of newsgroup documents posted on 20 different topics                                                                                                                                                      |
|                       | Sogou News     | Is a mixture of Sogou CA and Sogou CS news material, the classification label of news is determined by the domain name in the URL.                                                                                         |
|                       | Reuters        | Reuters News DataSet Reuters-21578 is the most widely used, and other data sets derived from a subset of Reuters data sets include R8, R52, RCV1 and RCV1-V2                                                                 |
| Smart question and answer | SQUAD          | The Stanford question-and-answer dataset is a collection of question-answer pairs derived from Wikipedia articles.                                                                                                       |
|                       | MS MARCO       | All questions are extracted from user queries and real web documents using Bing search engines, and answers are generally very creative and suitable for text generation                                                               |
|                       | TREC-QA        | There are two versions of the dataset, called TREC-6 and TREC-50. The former is divided into six categories of questions, while the latter is 20 categories for both versions, the training and test datasets contain 5,452 and 500 questions, respectively |
|                       | Quora          | Quora database is used for interpretation identification (detecting duplicate problems)                                                                                                                                    |
| Natural language inference | SNLI           | The Stanford Natural Language Reasoning (SNLI) dataset consists of 550, 152, 10,000, and 10,000 sentence pairs for training, development, and testing. Each pair is annotated using one of the three labels: Neutral, |
3. Text classification method based on depth blending model

After Kim proposed TextCNN model, more and more researchers are committed to transforming CNN network model to better complete a series of work from feature extraction to text classification. Some researchers combine some latest feature extraction and pretraining techniques with TextCNN model and apply them to practice. More researchers choose to mix CNN model with other neural network models, a new deep mixing model is constructed.

3.1 Improvements to the CNN model

With the introduction of the TEXT CNN model, the researchers began to focus on how to improve the CNN model to optimize the text classifier, and made some progress in this direction.

In 2017, Rie Johnson proposed a CNN structure that halves convolutional dimensions by fixing the number of convolutions: DPCNN [12], which effectively extracts remote relationship characteristics from text with less complexity.

J Zhou Liu presents the XML-CNN model [13] on the basis of TEXT CNN, which focuses on three improvements: using dynamic max pooling, i.e. for each feature map, pooling generates p characteristics; add a full connection layer between the pooling layer and the output layer. The XML-CNN model solves the problem of data sparseness in multi-label situations while the model training time is not slow.

Inspired by VGG and ResNets, Alexis Conneau proposed the VD CNN model [14], which has the same number of filters for layers of the same output feature map size after pooling operations, and if the size of the feature map is halved, the number of filters will double, greatly reducing memory consumption and increasing the training speed of the model.

3.2 A combination of CNN models and new technologies

To solve the problem that text CNN parameters are too large and generate word vectors inefficiently, Sun Ryan proposes a TextCNN algorithm [15] that combines Ada-BERT [16] and applies them to garbage bomb screen recognition and filtering: using ADABERT to learn bullet text to get more effective word vectors in less time; use batch standardization to reduce gradient loss. Experiments show that the algorithm improves the overall running speed of the algorithm and optimizes the performance on the screen data set.

Niu Yuxuan et al. put forward the ERNIE-DP CNN model based on DPCNN [17], which uses an enhanced language characterization model based on word vectors, ERNIE, to input the word vector output from the ERNIE pretraining model into the improved DPCNN model and form an improved ERNIE-DP CNN model for Chinese text classification experiments have shown that the model is superior to ERNIE-CNN in terms of data.

Currently, most publicly exposed text classification data sets are relatively balanced, but for real text distribution, categories are often extremely unbalanced, which can have an impact on model training. In response to this problem, Wang J L proposes a text classification algorithm based on convolutional neural networks and attention mechanisms-AM-CNN [18]. The algorithm uses the circular neural network to capture the context information of the text, and then uses the convolutional neural network model to complete the classification of the text by introducing the attention mechanism to obtain the
characteristic vector matrix of the text category, in order to deal with the extreme situation of some small categories in the data set.

Jian Hong MA et al proposed a text classification model (CGGA) [19] mixed by CNN and parallel gating mechanism in 2021, which entered the character-level word vector into the convolution layer to obtain the local feature matrix, which entered GTRU [20] [21], Bi-GRU [22] or context modeling and learned the connection between data contexts. The outputs of the two are then stitched together and entered into the Global Average Pooling layer (Global Average Pooling, GAP) [23] instead of the full connection layer for feature compression and prevention of overfitting. Then, the feature weight update calculation is carried out by using the multi-head attention mechanism [24] [25], so that the model has higher output in the corresponding text category. The model structure is shown in Figure 1:

![Figure 1 CGGA model structure diagram](image)

To verify the model superiority, the authors compared four neural network models on two data sets, THUC-News and SoGouCS, and the results showed that CGGA did perform better on the accuracy parameter than the other three hybrid models.

Table 2 Model accuracy comparison

| Model   | THUC-News MAP(%) | SoGouCS MAP(%) |
|---------|------------------|----------------|
| CCNN    | 93.93            | 89.33          |
| Bi-GRU  | 94.93            | 95.19          |
| CCNN-A  | 95.11            | 94.85          |
| CGGA    | 95.69            | 96.11          |

3.3 A mix of CNN models and other models

Considering CNN's advantage is that it extracts locally relevant characteristics of data, the disadvantage is that it ignores local contextual information, Wang Haitao, Song Wen, etc. take advantage of LSTM's ability to make better use of contextual information when dealing with sequence problems, and propose in 2020 a text classification model MLCNN (Merge-LSTM-CNN) [26] that combines long-term memory networks (LSTM) with convolutional neural networks. The model structure is shown in Figure 2:

![Figure 2 MLCNN model structure](image)

The model combines the advantages of LSTM's ability to capture contextual dependencies: first, vectorize input text by word embedding, extract local features of text through three layers of CNN, and
then integrate full-text semantics; at the same time, the characteristics of historical information in text are stored using LSTM, which captures the context-related semantics of text. In addition, the input vector is fused with the output of each layer of CNN to realize the reuse of the original features. To verify the superiority of the hybrid model performance and classification effect, the author compared the experiment on the THUC-News data set and evaluated the model on accuracy, recall rate and F1 value:

| Model      | Acc  | Rec  | F1   |
|------------|------|------|------|
| CNN        | 91.64% | 91.59% | 92.27% |
| LSTM       | 89.86% | 90.71% | 87.89% |
| SVM        | 94.33% | 94.09% | 93.38% |
| P-LSTM     | 92.05% | 93.81% | 93.53% |
| CNN-LSTM   | 94.07% | 93.14% | 95.16% |
| LSTM-CNN   | 95.42% | 95.19% | 94.95% |
| MLCNN      | 96.45% | 96.36% | 96.51% |

JinBao Teng et al. made improvements on its basis in 2021 by proposing a text classification method based on the LSTM-Attention hybrid model with CNN [28], which solves the problem that LSTM and CNN cannot reflect the importance of a single word in the overall text when extracting features. The model adds attention mechanism (Attention) to the LSTM to extract the attention value of the output information. Incorporating the output of LSTM-Attention with the output of CNN, the training results of the AG-news data set indicate that the hybrid model has a better text classification effect than other depth models. The specific data is shown in Table 4:

| Model      | Acc  | Pre  | Rec  | F1   |
|------------|------|------|------|------|
| CNN-LSTM   | 86.79 | 87.10 | 86.72 | 86.76 |
| Bi-LSTM    | 87.36 | 87.84 | 87.78 | 87.75 |
| Bert-Softmax | 86.62 | 86.93 | 86.78 | 86.89 |
| MLACNN     | 90.23 | 89.63 | 90.36 | 90.12 |

Li Ruihuan and others from the practical application, the MLA CNN model and criminal case judgment of the text classification scene combined, through the doc2vec text vector characteristics to optimize the hybrid model, put forward doc2vec-MLA CNN [29] hybrid model: according to the characteristics of the criminal sentence category uneven distribution and the proportion of the sentence category different using weighted loss function re-optimization, experimental data table proves doc2vec-MLACNN model does improve classification accuracy by being optimized twice.

All in all, most researchers are committed to improving the internal structure of the CNN model, or replacing the full connection with global average pooling, so as not to fit, or to replace the traditional convolution layer with a circular way and build a semi-pooled layer[30], and some scholars combine cutting-edge new technologies with deep hybrid neural networks, such as BERT technology, attention mechanism [30], etc., and some scholars directly combine traditional convolutional neural networks with LSTM, Bi-LSTM, etc. to complement each other and build a new deep hybrid model. No matter from what direction to the text classification model, its ultimate goal is to improve the efficiency of text classification, and then serve the reality, because there is no uniform text classification data set, and most of the models proposed by the researchers are only trained on their own designated data sets, it is difficult to compare the results of the researchers, which is also a concern of text classification researchers.
4. Summary and outlook
On the basis of the deep learning model, text classification with the times to join the attention mechanism, Transformer, BERT and other advanced ideas and technologies, in the text data for more detailed pre-processing at the same time, either the traditional neural network model to improve optimization, or some advanced technology and traditional neural network integration, or mixed with multiple neural network models, to make up for each other, in recent years rely on these technical methods text classification has indeed made good progress, but also exposed some areas that still need to be improved.

From the perspective of neural network model itself: most of the existing structures of shallow and deep learning models are tried to be used for text classification. Especially, the best model based on transformer can fine tune many NLP tasks by increasing data, improving computing power and designing training programs, therefore, how to make a direct choice without wasting computing resources and achieving high prediction performance will be the difficulty of future research.

From the aspect of performance evaluation: the existing shallow model and deep model can basically achieve good performance in most text classification tasks, but they lack the ability of anti-interference to the results. For example, when there are a small number of interference items in the data set, no matter the deep or shallow model cannot deal with them well, so improving the robustness of the model is also the current research hotspot.

From the perspective of data sets: Although many large-scale datasets of common text classification tasks have been collected in recent years, new data sets for more challenging text classification tasks still need to be trained, such as text classification for multilingual documents, text classification for large documents, etc.

From the perspective of transfer learning ability of the model [31]: Malte a and others put forward the application of transfer learning in natural language processing in 2019. Its main idea is summarized as follows: pretraining the main parameters of the model on general tasks, and pretraining the main parameters of the model on specific tasks, and only need to use small-scale data to retrain and fine tune the pretrained model. In this way, we can ensure that the model has a faster convergence speed and achieve better model effect. By combining transfer learning with the increasingly updated hybrid neural network, we can obtain higher experimental results and a more perfect model structure, also makes the model have the universal applicability brought by transfer learning.

Now, when there are complex and diverse text types, such as government platform comment text, news text, microblog community comment text, and other mixed types of text, how to achieve good results in different application fields will be determined by the transfer learning ability of the model.

In the era of rapid development of mobile Internet, text data on the Internet is growing. In the face of diversified text data, how to extract, express and classify features has become a research hotspot in the field of natural language processing, in particular, the paper analyzes and summarizes the direction of machine improvement of hybrid neural network text classification model which has been updated in recent years, hoping to provide useful thinking direction for scholars and technicians in related fields.

References
[1] Zha I S, Zhang Z M. (2016) Semi-supervised autoencoder for sentiment analysis. In: Thirtieth AAAI Conference on Artificial Intelligence. Shanghai. 1394-1400.
[2] Guan Q, Ye S, Yao G. (2009) Research and design of internet public opinion analysis system IITA International Conference on Services Science, Management and Engineering. In: IEEE. Zhangjiajie, 173-177.
[3] Li C, Chai Y M, Nan X F. (2016) Research on Problem Classification Method based on Deep Learning. Computer Science, 43 (12) : 115-119.
[4] Guo M C, Zhang Y. (2019) Gaussian Transformer: A Lightweight Approach for Natural Language Inference. In: Thirty-Third AAAI Conference on Artificial Intelligence. Hawaii. 191-219.
[5] Bengio Y, Ducharme R, Vincent P, et al. (2003) A neural probabilistic language model.
Journal of Machine Learning Research, 3: 1137-1155.

[6] Mikolov T, Sutskever I, Chen K, et al. (2013) Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems. In: NeurIPS, Nevada. 3111-3119.

[7] Mikolov T, Corrado G, Chen K, et al. (2013) Efficient Estimation of Word Representations in Vector Space. In: ICLR, Arizona. 1-12.

[8] G. Salton, A. Wong. (1975) A vector space model for automatic indexing. Communications of the ACM. 11.

[9] Zhou Z H. (2016). Machine Learning. Tsinghua university press, Nanjing.

[10] Kim Y. (2014) Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP. Doha. 1746-1751.

[11] LeCun, Yann, et al. (1998) Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, 11: 2278-2312.

[12] Rie Johnson. (2017) Deep Pyramid Convolutional Neural Networks for Text Categorization. In: Association for Computational Linguistics. Vancouver, Canada. 562-570

[13] Jing Zhou Liu. (2015). Deep Learning for Extreme Multi-label Text Classification. SIGIR:115-124

[14] Conneau A, Schwenk H, Barrault L, et al. (2017) Very Deep Convolutional Networks for Text Classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers.

[15] Jacob Devlin Ming-Wei Chang Ber. (2016) Bert : Bidirectional Encoder Representations from Transformers. https://arxiv.org/pdf/1810.04805

[16] Sun R A, Zhang Y H. (2021) TextCNN garbage barrage recognition and filtering algorithm combined with AdaBERT. Intelligent computer and application. 11(04):9-13.

[17] Niu Y T. (2021) Chinese text classification based on improved ernie-dpcnn model. Journal of Jiangsu Normal University (NATURAL SCIENCE EDITION). 39(01):47-52.

[18] Wang J L. (2019) AM-CNN : a Convolution Neural Network Architecture for Text Classification Based on Attention Mechanism. Minicomputer system. 40(04):710-714.

[19] MA J H, LIU Y P, LIU Y D. (2021). CGGA: Text Classification Model Based on CNN and Parallel Gating Mechanism. Journal of Chinese Computer Systems, 42(03):516-521.

[20] Zhang T. (2021-04-27) An improved text classification method based on TextCNN. Beijing: CN109918507B.

[21] Zhang T, Yang J. (2019). Research on the Improved Activation Function TReLU. Journal of Chinese Computer Systems. 40(1) : 58-63.

[22] Chao Zhou, Cai Hong Li. (2016) An Efficient Solution of Circuit State Assignment with Immune Algorithm. In: IEEE 6th International Conference on Electronics Information and Emergency Communication.

[23] Xue Wei, Li Tao. (2018) Aspect based sentiment analysis with gated convolutional networks. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL2018). Melbourne: 2514-2523.

[24] Cho K, Van Merrienboer B, Gulcehre C, et al. (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Conference on Empirical Methods in Natural Language Processing. 1724-1734.

[25] Shuo W, et al. (2019) Predicting EGFR mutation status in lung adenocarcinoma on computed tomography image using deep learning. European Respiratory Journal Mar. 53 (3) 1800986

[26] WANG H T, SONG W, WANG H. (2020) Text Classification Method Based on Hybrid Model of LSTM and CNN. Journal of Chinese Computer Systems.41(06):1163-1168.

[27] TENG J B, KONG W, TIAN Q X, WANG Z Q . (2021) Text Classification Method Based on LSTM-Attention and CNN Hybrid Model.
[28] Li R B. (2019). Research on Method of Predicting Prison Term in Criminal Cases Based on Mixed Deep Learning Model. https://kns.cnki.net/kcms/detail/detail.aspx?FileName=1020702793.nh&DbName=CMFD2020

[29] Wang J L, Peng D L, Chen Z, et al. (2019) AM-CNN: a convolution neural network architecture for text classification based on attention mechanism. Journal of Chinese Computer Systems. 40(4): 710-714.

[30] Ze H D, Wei D, Zhen H L, et al. (2019) Multi-task multi-head attention memory network for fine-grained sentiment analysis. In: Natural Language Processing and Chinese Computing, 8th CCF International Conference(NLPCC 2019), Dunhuang. 609-620.

[31] Malte A. (2019). Evolution of transfer learning in natural language processing Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. NAACL. Minneapolis. 15-18.