A Novel Optimized Convolutional Neural Network Based On Attention Pooling for Text Classification

Chengjie Yao\textsuperscript{1,}\textsuperscript{a} and Mengmeng Cai\textsuperscript{2,}\textsuperscript{b}

\textsuperscript{1,2}College of Electronic and Optical Engineering, Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu Province, China

\textsuperscript{a}email: 1019020919@njupt.edu.cn, \textsuperscript{b}email: caimengmeng1126@163.com.

Abstract. Convolutional neural network (CNN) is widely used in Natural Language Processing (NLP) and has achieved relatively good results. Since the convolutional layer of CNN can be initialized by encoding important semantic features, CNN can extract important semantic features in the convolutional layer. The pooling layer of CNN uses the feature of max-pooling to filter the features, but it does not consider the key information in the sentence and the context semantic information in the text. Therefore, a novel neural network algorithm combining with the attention mechanism and the pooling layer is proposed in this paper to make the model pay attention to the keywords in the sentence and automatically maintain the most meaningful text message, thereby improving the performance of text classification. Our paper uses important information features to initialize the convolution filter so that the model can notice the important semantic features. The resulting feature is enhanced by combining the attention mechanism in the pooling layer, so that the model can maintain the important information features of the text. Experiments show that the proposed model achieves excellent results for multiple text classification tasks, including sentiment classification and topic classification.

1. Introduction

NLP is a language that uses the computer to process and understand human language, and it has a wide range of applications. Text classification is the basic task of NLP, and it is applied to various fields of NLP, such as information retrieval, information filtering, and semantic analysis [1]. With the rapid development of neural networks, people are accustomed to processing natural languages with Recurrent Neural Network (RNN), CNN, or a combination of RNN and CNN [2].

[3] achieved the sentence-level simple classification by using CNN for feature extraction, which fully reflects the superiority of CNN in natural language processing. In general, when training a neural network, the effect of word vectors is related to the size of the corpus, which means it requires pre-trained word embedding to increase the corpus of processing tasks to support the experiment. The attention mechanism is an important mechanism for obtaining useful information in neural networks. [5] proposed a general attention-based convolutional neural network to model sentences. [6] introduced a self-attention mechanism to extract embedding models of interpretable sentences. [7] proposed a classifier based on the attention mechanism, which can predict various emotions in different sentences. In [8], the application of attention mechanism to convolutional neural networks has also been successful. A novel neural network architecture combining attention mechanism and max pool was proposed to automatically maintain the most meaningful information of the text by using a bidirectional gating recursive unit (BGRU) [9]. The pooling layer is an important part of the convolutional neural network. The max-pooling layer is often used to filter important features, but when the important information is
distributed in different parts of the sentence, the max pooling may fail. Therefore, this paper considers
the combination of attention mechanism and pooling operation to maintain important text information.

Based on the above references, this paper proposes a novel text classification method based on the
initialization convolution filter, which combines the attention mechanism and the neural network
algorithm of the pooling layer. We initialize the weights by encoding the semantic information into a
filter to filter out important information before training the neural network. Then we use the attention
pool-based strategy to enhance important information to retain the most valuable feature information.
By using the operation of the attention pool, we have achieved better performance on the tasks of
emotion classification and topic classification.

2. Model
As is shown in Figure 1, our model is mainly composed of four parts, including an input layer, a
convolutional layer, a pooling layer and an output layer. The model first uses CNN with initialization
weights so that the model can focus on the important information of the text before training. Because
the important information of the sentence is distributed in different regions, the model does not directly
maximize the output of the convolution layer, but uses an attention-based pooling scheme to solve the
pooling problem. This model not only helps the convolution layer filter to filter out more important
features, but also helps prevent filtering out important classification features that has not been captured
by CNN with the feature enhancement operation in the pooling layer.

![Figure 1. Process of text classification based on attention pooling](image)

2.1. Input layer
In our model, the input is a sentence $S$. Each word $i$ in the sentence $S$ has a corresponding word vector
$x_i \in R^k$ ($k$ represents the number of dimensions of each word vector), and each word is embedded as a
training word. Let $a_i \in A^m$ be the d-dimensions vector of the word $x_i$, and $C \in R^{l \times d}$ represent the word
embedding matrix, where $l$ is the maximum embedded sentence length.
2.2. **Convolutional layer**

Since CNN can easily capture the semantic features of n-gram, our paper uses important n-gram to initialize filters, which allows neural networks to identify important semantic information before training, so that the effective position information of each sentence can be captured in the convolutional layer. For binary classification, the formula for calculating the weight $w$ according to the Naive Bayes algorithm is:

$$w = \frac{(p+\alpha)/||p||_1}{(q+\alpha)/||q||_1}$$

where $\alpha$ is a smoothing parameter, $p$ is the number of texts that contain n-gram in class $i$, and $q$ is the number of texts contain n-gram in other classes. $||p||_1$ and $||q||_1$ represent the number of $i$ and other classes in the text, respectively.

![Figure 2. Initialization of filter size 3, 4, 5](image)

Figure 2 shows that how we fill uni, bi and tri-gram feature into the filters. By initializing the important positions of the CNN, a new convolution layer weight $W_n$ is obtained, and the output of the convolution layer is

$$O_n = \tanh(W_nx_i + b)$$

where $x_i$ is the input of the embedding layer, $b$ is the bias, and the activation function is the nonlinear hyperbolic tangent transformation function $\tanh$.

2.3. **Attention pooling**

As shown in Figure 2, the attention mechanism generates matrix $A^m$, which is formed by obtaining the weight of each phrase from the feature vector of the convolution layer, that is,

$$H_a = \tanh(W_1O_n)$$

Because the importance of different words in a sentence may also be different, our paper uses a softmax function to process the model matrix $H_a$ to obtain the attention pool matrix $A^m$, the formula is as follows,

$$A^m_{i,j} = \frac{\exp(H_{i,j})}{\sum_{j=1}^{n}\exp(H_{i,j})}$$

where $H_{i,j}$ is a $(i,j)$ dimensional matrix, and $A^m_{i,j}$ is also a $(i,j)$-dimensional matrix. Finally, we multiple the obtained attention pool matrix and the convolution output $O_n$, and highlight the importance of certain words. Then we enhance the resulting features of the output of the convolution layer. The formula is as follows.

$$M_a = O_nA^m$$

where $O_n$ is the output vector of the convolution layer. $W_1$ is the parameter matrix, and $M_a$ is the output of the pooling layer.

2.4. **Output layer**

For text classification, our paper divides sentences into positive and negative categories according to different tasks. The softmax function calculates the probability score of each category to predict the category to which the text belongs. The formula is as follows.

$$\hat{y} = \text{softmax}(W_3M_a + b)$$

where $\hat{y}$ represents the probability of the category to which the text belongs, $W_3$ and $b$ are the weight matrix and bias respectively. The cross-entropy function is defined as

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}(y_i \log(\hat{y}))$$
where $y_i$ and $\hat{y}_i$ are the true value and the predicted value respectively, $m$ is the number of texts.

3. Experiment

3.1. Experimental setup and dataset

Our paper uses Google News to pre-train word embedding layer through the word2vec toolkit, and uses uni, bi, and tri-gram features to initialize filters. In order to prove the accuracy and effectiveness of the model, this paper conducted experiments on five datasets based on tensorflow, and compared our model with the latest model. On the initialization convolution filter, this paper uses k-means algorithm to cluster the n-gram initialization filter. We test our model on five benchmark datasets for sentence classification tasks, which are the same as those used in [4]. The statistical information is listed in Table 1. Table 2 is our experimental parameters.

| Dataset | Classes | S  | Train  | P   |
|---------|---------|----|--------|-----|
| MPQA    | 2       | 36 | 10605  | 6083|
| SUBJ    | 2       | 120| 10000  | 17896|
| MR      | 2       | 63 | 10662  | 16448|
| SST2    | 2       | 52 | 6920   | 12736|
| CR      | 2       | 105| 3774   | 5046 |

Table 1. The summary statistics of benchmark datasets

| Parameters         | Setting |
|--------------------|---------|
| Filter size        | 3,4,5   |
| Learning rate      | 1       |
| Min batch size     | 12      |
| Dropout            | 0.5     |
| Iterations         | 10      |

Table 2. Experimental parameters

3.2. Experimental results and analysis

It should be noted that the data extracted by the initial convolution filter may have an impact on the pooling layer. The author is inspired by this. In order to further explore the impact of different n-grams on the model, setting performance indicators is accurate. Take the datasets MPQA and SUBJ as examples, as shown in Figure 3. Experiments show that different n-gram initialization filters produce different results at different iteration times. The number of iterations corresponding to the optimal performance of 1-gram, 2-gram and 3-gram are also different. For the dataset MPQA, while 1-gram initializes the filter, when epoch=12, the performance reaches the optimal value of 90.4%; when epoch=10, the accuracy of the 2-gram and 3-gram initialization models reaches the optimal values, which are 89.9% and 89.8%, respectively. When the dataset SUBJ trains the model, the optimal values of 1, 2 and 3-gram initializing the model are 94.1%, 94.2% and 93.9% respectively, and their corresponding iteration times are 12, 14 and 12, respectively.

![Figure 3. Accuracy of the training set of MPQA and SUBJ at different epochs](image)
Figure 4. The loss value at different epochs

Table 3 Accuracy of the model of Max Pooling-based CNN (MPCNN) and Attention Pooling-based CNN (APCNN) at different n-gram initialization

| Model      | MPQA  | MR   | SST2 | SUBJ | CR  |
|------------|-------|------|------|------|-----|
| MPCNN+UNI  | 89.3  | 82.1 | 89.0 | 93.7 | **86.0** |
| APCNN+UNI  | **90.4** | **82.9** | **90.3** | **94.3** | 85.9 |
| MPCNN+BI   | 89.5  | 82.2 | 88.3 | 93.7 | 85.8 |
| APCNN+BI   | **89.9** | **82.8** | **91.2** | **94.0** | **86.1** |
| MPCNN+TRI  | 89.2  | 82.1 | 88.2 | 93.8 | **85.9** |
| APCNN+TRI  | **89.8** | **82.4** | **90.5** | **94.2** | 85.8 |

In order to analyze the results comprehensively, the author conducted experiments on the convergence of the training set. We compared the datasets MR and MPQA of different n-gram initialization filters to better present the results. The number of iterations is set to 16, and one point is recorded every two iterations. The training results are shown in Figure 4. The following observations can be learned from Figure 4: When 1, 2, and 3-gram initialize the model on MR, the initial value of training loss is about 0.45. After a certain number of iterations, the convergence speed of the model initialized by 1-gram is significantly higher than that of 2-gram and 3-gram. Similarly, for the dataset SUBJ, as the number of iterations increases, the convergence rate of the 1-gram initialization model is higher than that of the 2-gram and 3-gram.

Table 4 Classification accuracy results of other models on benchmark datasets

| Model       | MPQA  | MR   | SST2 | SUBJ | CR  |
|-------------|-------|------|------|------|-----|
| CNN-non-static | 89.5  | 81.5 | 87.2 | 93.4 | 84.3 |
| MP-CNN(UNI) | 89.3  | 82.1 | 89.0 | 93.7 | 86.0 |
| MV-CNN      | -     | -    | 89.4 | 93.9 | -   |
| MGN-CNN     | -     | -    | 88.3 | 94.1 | -   |
| DSCNN       | -     | 81.5 | 89.1 | 93.2 | -   |
| Adasent     | 83.1  | -    | -    | 95.5 | **86.3** |
| AP-CNN (UNI)| **90.4** | **83.1** | **90.3** | 94.3 | 85.9 |

Experimental results show that for the same dataset, the convergence rates of different n-grams to initialize the model are different. And we believe that when 1-gram initializes the filter, the model converges faster.

Through the comparison of the results in Figure 4, we know that different n-gram initialization filters achieve different optimal values. For different datasets, although the convergence speed of 1-gram initialization model is faster, its corresponding accuracy is not always the highest. Therefore, when
choosing different gram to initialize the convolution filter, we must consider not only the impact on the training accuracy, but also the convergence speed.

Table 3 compares the training accuracy of the model proposed in this paper and in [4]. The experimental results prove that for different gram initialization convolution filters, attention pooling based CNN our proposed training accuracy is better than max-pooling CNN on most datasets except for the dataset CR. Both models use the same CNN filter initialization method. The model our proposed adopts a new attention pooling scheme to collect information features extracted by CNN. The scheme extracts the most significant information contained in the sentence through the attention weights. The latter only uses the max-pooling scheme to filter the dataset extracted by CNN. The experiment further shows that the attention pooling strategy is able to reserve more information contained in the sentence.

Table 4 shows the experiments in all tasks, we can conclude that the model we proposed outperforms the other systems except the datasets SUBJ and CR. Using n-gram to initialize the convolutional layer allows CNN to obtain better generalization ability. Based on the effective information features extracted by the improved CNN, the attention pooling can extract the most significant information contained in the sentence.

We believe that on the basis of filter-initialized CNN, the combined attention pool strategy is better than other CNN maximum pooling strategies. It not only because CNN can obtain better generalization ability, but also because the attention pooling can assign greater weight to more important information features. Therefore, this merger strategy has obvious advantages. The effectiveness of the proposed model has been confirmed in experiments.

4. Conclusion
This paper proposes an optimized CNN based on attention pooling, namely CNN-Att-pooling. In previous CNN based on semantic feature initialization weights, a novel attention-based pooling layer is also combined. In the text classification process, not only the important semantic features can be filtered at the convolution layer, but also the important semantic features can be enhanced by combining the attention mechanism at the pooling layer, thereby effectively improving the performance of text classification. The experimental results confirm the effectiveness of the proposed model. However, the CNN filter extract n-gram features without considering the relationship between sentences, so it still needs further research to apply CNN to the NLP fields.

Acknowledgements
This research was supported by the Natural Science Foundation of China (No.61977039), “Study of Mechanism and Classification for EEG Emotional Brain network in Teaching Situation”.

References
[1] Aggarwal, C.C., and Zhai, C. 2012. A survey of text classification algorithms. In Mining text data, pages 163-222.
[2] P. Liu, X. Qiu, and X. Huang. Recurrent neural network for text classification with multi-task learning. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016, pages 2873–2879.
[3] Yoon Kim. Convolutional neural networks for sentence classification. In Proceedings of EMNLP 2014, pages 1746–1751.
[4] Shen Li, Zhe Zhao, Tao Liu, Renfen Hu, and Xiaoyong Du. 2017. Initializing convolutional filters with semantic features for text classification. In Proceedings of the 2017 Conference on Empirical Methods.
[5] Wenpeng Yin, Hinrich Schutze, Bing Xiang, and Bowen Zhou. 2015. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. arXiv preprint arXiv:1512.05193.
[6] Zhouhan Lin, Minwei Feng, C’icero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou,
and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In ICLR.

[7] Y. Kim, H. Lee, and K. Jung. 2018. Attnconvnet at semeval-2018 task 1: Attention based convolutional neural networks for multi-label emotion classification. CoRR, abs/1804.00831.

[8] J. Du, L. Gui, R. Xu, and Y. He. 2017. A convolutional attention model for text classification, in Proc. Nat. CCF Conf. Natural Lang. Process. Chin. Comput. Cham, Switzerland: Springer, pages 183–195.

[9] Dejun Zhang, Mingbo Hong, Lu Zou, Fei Han, Fazhi He, Zhigang Tu, Yafeng Ren. 2019. Attention Pooling-Based Bidirectional Gated Recurrent Units Model for Sentimental Classification. In International Journal of Computational Intelligence Systems, pages 723 – 732.