Text Classification using Triplet Capsule Networks

Yujia Wu[1], Jing Li*[1], Vincent Chen[2], Jun Chang[1], Zhiquan Ding[3], Zhi Wang[3]
[1]School of Computer Science, Wuhan University, Wuhan, 430072, China
[2]Pembroke College, Oxford University, Oxford OX1 1DW, UK
[3]Sichuan Institute of Aerospace Electronic Equipment,Chengdu, 610100, China
{wuyujia,leejingcn}@whu.edu.cn, vincent.chen@pmb.ox.ac.uk, chang.jun@whu.edu.cn, {252241227,934877270}@qq.com

***Abstract*—Most existing methods only consider the local features of the samples, and their experimental results show better performance than traditional Non-deep learning methods. However, in these methods, the global features of the sample space are usually ignored, and these ignored global features will affect the classification accuracy. To solve this problem, a novel triple capsule network framework is proposed to text classification. The training in the first stage, to obtain a basic capsule network for obtaining local features. Then, three capsule networks sharing parameters are combined spatially, and the triplet loss function is used in the second stage of training. By comparative learning, the capsule network can learn global features that can represent the spatial distance between different categories. Through comparison experiments on six datasets and ten general benchmark algorithms, the results show that our results is the first in the four datasets.**

***Index Terms*—deep learning, text classification, capsule network, triplet loss**

I. Introduction

Natural language processing (NLP) is an important research area of current Artificial Intelligence technology, which includes Text classification [1], Text Clustering [2], Network Computing [3], Personalized Recommendation [4], Question Answering [5], Learning Semantic Representations [6], and so on. Convolutional Neural Network(CNN) [7] and Recurrent Neural Network(RNN) [8] both achieved good experimental results. The RNN extracted text features of contextual relationships, and CNN algorithm by convolution operation is used for exacted the local features on the samples. Then use the maximum or average pooling operation and connect to the output layer through a fully connected layer.

In this process, the spatial positions of features of the sentence are not considered. To address this problem, capsule networks can use vectors instead of neurons [9]. The advantages of the capsule network allowed scholars to use the information aggregation scheme for the capsule network model, which achieved good results of five text classification [10]. The capsule network [11]–[13] has achieved very good application results in the sentiment classification problems [14] and transfer learning problems [15].

However, the capsule network has achieved better than CNN and other models in the field of text classification, there are still some challenges. First, the processing object of this model is a single sample of a certain category. In this process, through supervised training of labeled samples, the global semantics of different categories of the entire dataset are not considered. Secondly, they usually ignore the information representing the global categories space distance between samples, and these global features include the global categories differences between the samples.

Therefore, it is important to obtain global features that can represent the global semantics of each category and the global distance information on the global category space. For example, some features frequently appear in various categories in the text dataset; they are called global shared features, and it is difficult to effectively distinguish them by the existing methods. As shown in Figure 1, the red dotted box represents the features belonging to category A, and the green dotted box represent the features belonging to category B. However, some global shared features are difficult to distinguish from their specific categories, such as features at the intersection of two dashed boxes, which may belong to category A or B. Because we do not have enough information to distinguish these, global shared features.

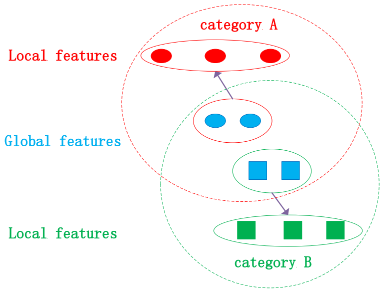

Fig. 1. Differences between some global shared features of categories A and B.

For this problem, we built a triple capsule network framework. First, we constructed a basic capsule network that uses capsule vectors instead of neurons. Then, in the first stage, we use labeled text dataset to train the basic capsule network. After completing the training, the three basic capsule networks sharing the weight parameters are combination of the second-stage training through the triple loss function. In this training process, for each sample, a sample of the same type is randomly selected as a positive sample, and a different

Corresponding author: Jing Li. E-mail:leejingcn@whu.edu.cn.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

sample is randomly selected as a negative sample to form a triplet sample. Through comparative learning, learning global features that can represent global semantics. The training goal is to make the spatial distance between samples of the same categorys as small as possible, and for samples of different categorys, increase as much as possible. Through the training in the second stage, the basic capsule network can learn the global category differences between these samples and effectively distinguish some difficultly distinguished global shared features. For example, in Figure 1, we need to effectively classify the global shared features of the intersection of the two dashed boxes to determine whether they belong to category A or B. This study applies the triplet loss to text classification. Through comparative learning, the neural network can learn global features that represent global differences. Compared with other benchmark algorithms, our method shows significant improvement. The main innovations of the article are as follows:

- This study applies the triplet loss to text classification. Through comparative learning, the neural network can learn global features that represent global differences.
- Previous methods only considered local feature information. In contrast, our method further considers local and global features and trains through two stages.
- The benchmark dataset is validated by our proposed method, which is superior to some benchmark algorithms.

II. RELATED WORK

Traditional machine learning methods [16]–[21] are not as effective as some deep learning-based models [22], [23]. At present, when it is to deal with text classification problems, it is mainly some deep learning such as CNN networks [24], [25], RNN networks [26], Generative Adversarial Nets [27], and Attention Networks [28]–[31]. Wang proposed to integrate the convolution operation into the RNN model [32]. Howard, et al. [33] proposed a general language model. Zeng, et al. [34] proposed a text classification model for solving the sparseness of short text data. Deep Pyramid CNN can obtain the correlation between features and improve the performance of the model to some extent [35]. Zhang, et al. [36] proposed to use CNN algorithm to process character-level text classification tasks. Kalchbrenner, et al. [37] uses a dynamic model for merging sentence lengths to improve the applicability of the model. Word vectors to represent words of representation spaces and are an important step in NLP's processing of natural language [38]–[40]. On this basis, Wang et al. [41] proposed the label joint word vector representation method of text classification and achieved good results.

However, in some cases, the CNN model has some limitations. Hinton et al. uses capsule networks to solve this problem [42], [43]. Xi et al. [44] studied the potential for the capsule networks on the CIFAR10 dataset. Jaiswal et al. introduce capsules into GAN and achieved good results [45]. Verma et al. [46] studied graph capsule networks. Capsule networks have also gained good applications for object segmentation [47] and Cross-domain sentiment classification [48]. Xiao et al. [49] proposed a model based on capsule networks, using the advantages of capsules for feature clustering. The advantages of the capsule network are also applied to the legal field. Chalkidis et al. apply the capsule network to solve the classification problem in the legal field [50], and used various preprocessing tools to make the capsule network adapt to research tasks of different backgrounds. Liu et al. propose a generative interpretation model that can learn to make classification decisions and simultaneously generate fine-grained interpretations [51]. Gururangan et al. [52] proposed a lightweight pre-training framework that can perform effective text classification when data and computing resources are limited.

III. PROPOSED MODEL

We use a triplet capsule network of processing text classification tasks. In this section, first of all, the formalized problem, then introduced the proposed network structure and triplet loss function, and finally introduce the basic capsule network structure.

A. Problem Formalization

A dataset comprises documents of categories C , a document of a category $D^c = \{T_1, T_2, \cdots, T_k\}$ contains sentences, and each sentence T_k=$\{i_1, i_2, \cdots, i_m\}$ contains m words. In the triplet capsule network framework, the input of the network is a triplet sample T_k^t=$\{a_k, p_k, n_k\}$, which includes three samples, an anchor sample a_k, and a label y_k. A positive p_k sample is randomly selected from the same class of samples as a_k. Another negative sample n_k is selected from a type different from that of sample a_k. In the triple capsule network framework, three basic capsules networks share parameters.

The training goal of the second stage of the triplet capsule network is to reduce the distance between the anchor and the positive samples as much as possible, and increase the distance between the anchor and negative sample as much as possible, which uses a triplet loss as the objective function.

B. Triplet Loss

The triplet network was first proposed by Hoffer et al. [53], and its purpose is to learn useful global representations of data by comparing distances. In this article, the triple capsule network comprises three basic capsule networks. The input of the network is a triplet sample T_k^t=$\{a_k, p_k, n_k\}$, where a_k and p_k are samples from the same class, and from other classes. Among them, the basic capsule networks map a variable-length texts sequence of a set of capsules of a fixed size.

In Figure 2, the training goal of the triplet capsules network is to reduce the distances between the two capsule groups of the same sample as much as possible, such as the capsules produced by a_k and p_k; and to increase the distances between the two groups of capsules from different categories as much as possible, such as the resulting capsules. Construct a loss functions to accomplish this. However, it is a problem here. The feature produced by the traditional CNN or RNN is a set of scalars, that is, a vector. The two vectors can easily

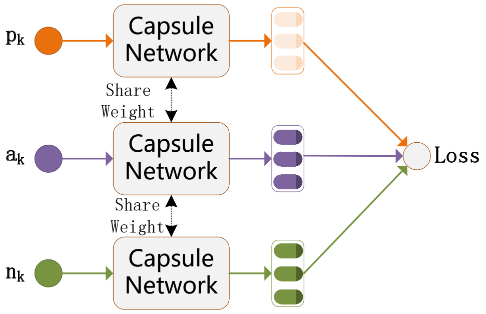

Fig. 2. Triplet loss structure. Input a triplet sample and obtain global features of comparative learning.

compare the distance. For example, the cosine distance can be used to represent the distance well. However, the triple capsules network generates a set of capsules, and each capsule is a vector. It is not possible to directly measure the distance between the two groups of capsules by using the cosine distance and other methods.

To solve this problem, given two sets of capsules, our method will have a predefined capsule similarity function. Let, Cap_1, Cap_2 be respectively expressed as two sets of capsules, with the length of each capsule equal to 8. Cap set of capsules can be represented as a matrix of size $m \times 8$. The definition of the distance $R(Cap_1, Cap_2)$ between the two groups of capsules Cap_1 and Cap_2 is

$$R(Cap_1, Cap_2) = \frac{\langle Cap_1, Cap_2 \rangle}{\|Cap_1\| \, \|Cap_2\|} \tag{1}$$

The matrix norm $\|\cdot\|$ is expressed as an equation, for example, $\|Cap\| = \sqrt{\langle Cap, Cap \rangle}$. $\langle Cap_1, Cap_2 \rangle$ is the inner product of matrices Cap_1 and Cap_2 and is calculated as

$$\langle Cap_1, Cap_2 \rangle = TR({Cap_2}^T Cap_1) \tag{2}$$

In formula (2), ${Cap_2}^T$ is the transpose of matrix Cap_2, and at the same time, $TR(\cdot)$ is to accumulate the diagonal element.

When $R(Cap_1, Cap_2) = 0$,the distance between the two is the maximum. At this time, the two groups of capsules have the worst similarity. When $R(Cap_1, Cap_2) = 1$, the two capsules have the shortest distance, and the two capsules have the best similarity. The loss function is

$$Loss = \sum_{i}^{N} [\|f(a_k) - f(p_k)\|_2^2 - \|f(a_k) - f(n_k)\|_2^2 + \alpha] \tag{3}$$

where α is the distance limited to be maintained between the same and different categories. Here, the model learned by comparing samples, different from other neural networks in the way of training through labels. Through comparative learning, the model learned the global difference between each of these documents, and these difference informations represents the global features. Therefore, after the first stage of training is completed, we use the triplet network of the second stage of training.

C. Local Feature Extraction

We previously described the triple capsule network framework of Section III-B. In this framework, three capsule networks of shared parameters are used for spatial cascade to learn global features. Here, we explain in detail how basic capsule networks work. Its goal is to learn a mapping that maps a variable-length text sequence in a fixed-size set of capsules (regard them as two-dimensional vectors).

Basic capsule networks use supervised training for labeled y_k samples T_k=$\{a_k\}$ as input to the model when trained separately. The input of the model is $T_{\rm d}(1 \leq {\rm d} \leq n)$ in $D = \{T_1, T_2, \cdots, T_n\}$ of a certain category. Each sentence T_d=$\{i_1, i_2, \cdots, i_k\}$ contains k words. The last layer of the model corresponds to the probability of each categorie. Figure 3 depicts the basic structure of basic capsule networks.

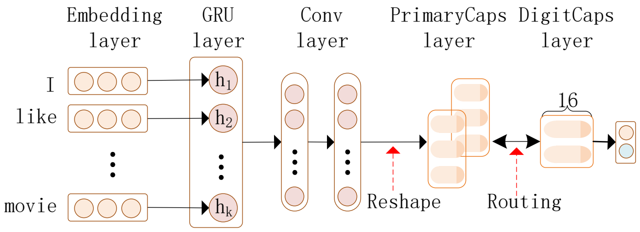

Fig. 3. Basic capsule network for local feature extraction.

In Figure 3, the input of the model is a variable-length text T_d, and the sentence T_d is passed to the GRU layer after word embedding representation, which is used to extract the features in the sentence T_d that have a context relationship between adjacent words. This is shown below.

$$h_k = [GRU(h_{k-1}^f, i_k); GRU(h_{k-1}^b, i_k)] \tag{4}$$

Here, the GRU uses a gating mechanism to make the RNN remember the past information, and also selectively filter some unimportant information. GRU layers extract local features of context information by extracting forward and backward text features. The input sentence $T_{\rm d}$ is subjected to feature extraction at the GRU layer, Composition feature map H, it contains the context extracted by the bidirectional GRU encoder, as shown below.

$$H = [h_1, h_2, \cdots, h_k] \tag{5}$$

In order to further extract more local features, we imitated the scheme for Sabour et al. [9]. Here, we adopted two convolutional layers, where the size of the Conv1 layer is 256 ,Window size is 3×3, stride setting 1, and ReLU activation. The Conv2 layer also has 256 3×3 convolution kernels. To obtain more information, set its stride to 2 and ReLU activation.

The fifth layer of the capsule network are the primary caps layer, which is a reshape operation of the Conv2 layer. Its purpose is to combine eight adjacent high-level features of an 8D capsule vector. Capsules can save the positional relationship of adjacent features. Here, the capsule may save the grammatical structure information and spatial distance of local features. The sixth layer of the model have a digital capsule layer representing the category.There are as many categories as capsules. As in the study of Sabour et al. [9], the dimensions of the digit caps layer capsules are set to 16D.

In most cases, CNN will use a pooling operation,this is a simple and efficient way for aggregating information. However, it is possible to lose considerable information. The capsule network has abandoned the pooling operation, and it performs a mapping between the digital capsule layer and the primary caps layer, using a routing algorithm. The following briefly introduces the process of the routing algorithm.

First, set a variable $b_{ij} = 0$, and then calculate it using

$$c_{ij} = \frac{e^{b_{ij}}}{\sum\limits_{k} e^{b_{ik}}} \tag{6}$$

In the initial state, the coefficient c_{ij} is equal to $\frac{1}{k}$, which implies that the next capsule is the weighted sum of each capsule in the previous layer, and the initial weight is $\frac{1}{k}$. The goal of the routing algorithm is to find the most appropriate weight coefficient.

After coefficient c_{ij} is determined, then iterate once using the following equations to obtain a new coefficient b_{ij}. At this point, the routing process is complete. For details, refer to the original article [9].

$$s_j = \sum_i c_{ij} W_{ij} u_i \tag{7}$$

$$b_{ij} = b_{ij}+W_{ij}u_i \cdot \frac{\|s_j\|^2}{1+\|s_j\|^2}\frac{s_j}{\|s_j\|} \tag{8}$$

where W_{ij} is a fixed shared weight matrix. Generally, After 3 or 4 iterations, the capsule network can achieve the best performance, and finally use the modulus length of the digital capsule to express the probability. We applied basic capsule networks of dynamic routing to text classification and obtained better results than previous algorithms such as CNN.

IV. EXPERIMENT

To verify our proposed model, six common publicly available benchmark datasets were used. Experiments were also conducted with ten benchmark algorithms. In these methods, 300-dimensional word vectors are used for preprocessing.

A. Datasets

To test our proposed model, in the experiments, we selected six benchmark datasets: Movie Review [54], Multiple-Perspective QA [55], Subjectivity Datasets [56], Stanford Sentiment Treebank 1 [57], Stanford Sentiment Treebank 2 [57], TRECQA [58]. The statistics of the dataset are summarized in Table I. Movie Review is abbreviated as MR, Multiple-Perspective QA is abbreviated as QA, SUBJDA is abbreviated as SU, Stanford Sentiment Treebank 1 is abbreviated as S1, Stanford Sentiment Treebank 2 is abbreviated as S2 , TRECQA is abbreviated as TR.

TABLE I
SUMMARY STATISTICS FOR THE DATASETS

Dataset	Class	Length	Size	Vocabulary	Test
MR	2	20	10662	18765	2132
QA	2	3	10604	6246	2120
SU	2	23	9999	21323	2000
S1	5	18	11855	17836	2210
S2	2	19	9613	16185	1821
TR	6	10	5891	9592	1178

In Table I, class indicates the number of target classes, length represents the average sentence length, size indicates the size of the dataset, vocabulary represents the vocabulary contained in the dataset, and test is the size of the test dataset that was set.

The following computer was used in the experiment: 3.2GHz i7-8700 CPU; 11G GPU; 32G memory; Operating system is win10. The algorithm was implemented under Tensorflow 1.10 framework. For training, we first preprocessed from Word2vec using a word2vec vector to initialize the embedding vector. Choose 10% of the data onto testing. The batches of size 128. We used the Adam optimizer with an initial learning rate of 0.0003. The dropout regularization are 0.2. In our proposed model, the convolution kernel size of both convolutional layers was 3, and the number was 256.

B. Baselines

To verify the performance of the proposed algorithm, we selected ten latest benchmark algorithms for comparison including:

CNN: This model uses a pooling operation and a convolution operation to learn local features from the samples and finally get the output. It is the first deep neural network model used for text classification [7].

LSTM: LSTM is a RNN that acquires long-term and short-term text sequence information to improve text classification accuracy [59].

RCNN: A model that combines CNN and RNN. By using convolutional layers to extract features, the maximum pool layer is used to automatically determine which words play a key role in text classification to capture key information in the text [26].

RNN: By using a recursive structure, it is specific to multitasking,it can get the contextual information on the text [8].

HAN: A model based on attention mechanism not only improves the interpretability of the model, but also improves the accuracy of text classification [29].

BLSTM: Apply the obtained 2D maximum merge operation to improve the accuracy of text classification [60].

ULMFiT: General language fine-tuning model, it has very good performance in text classification tasks [33].

DR-AGG: It is the basic version of the capsule network used in text classification [10].

Capsule: Available in two different types of matrix capsule

networks for text classification [11].
USE: The full use of sentence vector representation of general encoder can improve the accuracy of text classification [61].

C. Overall Performance

The evaluation metric was classification accuracy. The overall accuracy of the proposed and benchmark algorithms on six datasets is listed in Table II.

Our proposed method shows that the best performance were achieved on all five datasets, and the S2 and TR also showed almost optimal results. In Table II, bold red indicates the result of the first place, and black bold and underlined indicates the result of the second place.

TABLE II
EXPERIMENTAL RESULTS

Method	***MR***	***QA***	***SU***	***S1***	***S2***	***TR***
CNN	81	89.1	93	42.3	86.8	92.8
LSTM	76.3	87.7	82.6	43.2	79.9	89.3
RCNN	81	88.3	90.3	44.1	82.9	90.7
RNN	80.6	88.6	89.4	43.8	80.6	90.8
HAN	77.1	87.4	89.1	47.2	83.5	87.1
BLSTM	82.3	89.1	**94**	50.4	**89.5**	96.1
ULMFiT	82.1	88.9	93.2	50.3	89.3	95.8
DR-AGG	82.4	88.9	93.1	**50.5**	87.6	92.4
Capsule	81.3	89.2	93.3	50.1	86.4	91.8
USE	81.5	86.8	93.3	50.3	87.2	**97.4**
LCaps (ours)	**83**	**89.2**	93.8	50.2	88.7	96.1
TriCaps (ours)	**83.1**	**90.3**	**94.2**	**50.6**	**89.3**	**96.2**

Among them, LCaps is our benchmark capsule network, which includes a GRU layer for extracting context information and two convolutional layers. It is superior to the traditional CNN algorithm on six datasets. However, LCaps did not consider global features, did not learn global class difference information, and failed to achieve the best performance. The TriCaps model, which considers global features, has better experimental results of six datasets than LCaps. It can be seen that the acquisition of global features can improve the classification performance. The TriCaps model, which combines global and local features, achieves the optimal result of all four datasets.

For the MR dataset, each review divided into two categories. On the MR dataset, the LCaps algorithm achieved results that exceed some of the benchmark algorithms. This may because the added GRU unit extracted the context information, which is helpful for improving classification accuracy. On such movie review datasets, models with recurrent structure perform well. For example, BLSTM can achieve accuracy of 82.3%. The TriCaps models consider global features have improved the basic performance of the LCaps model to a certain extent.

The two datasets of QA and SU are also two-category datasets. The QA dataset uses opinion tendency as a classification task, including positive and negative categories. SU is a subjective dataset, which is divided into two categories according to whether the sentence is subjective or objective. LCaps is comparable to the benchmark algorithm for these two datasets, while TriCaps models with global features are better than the benchmark algorithm. The maximum sample length of the SU dataset is 121, and most samples are relatively long. This helps to obtain better global features, so for this dataset, the TriCaps model that combines local and global features performs best.

S1 is a movie review dataset; S2 is based on S1, which removes neutral reviews and divides reviews of two categories, positive and negative. For both types of datasets, BLSTM and ULMFiT have shown very good results, especially for the S2 dataset, where the classification accuracy of BLSTM reached 89.5%. For the S1 dataset, TriCaps achieved the best results, and for the S2 dataset, TriCaps achieved the second-best result.

TR is a problem dataset, which is used to classify problems that are divided into six categories. The USE model obtained the best results, and the accuracy of classification was 97.4%, primarily because it used Google's corpus to learn the global sentence embedding representation, making it more capable of classifying sentences. In addition, the TR dataset divides questions on six types. Many words often appear in the sample, such as "how","what","when"; Thus, they may interfere with other global characteristics, leading to the outcome of the second place TriCaps.

In the above experimental analysis, TriCaps outperforms the benchmark algorithm on four datasets. On some datasets, it performs best result of some datasets. Although TriCaps algorithm may still have shortcomings. But TriCaps achieved the best results of all four datasets.

D. Impact of Training Size

In terms of accuracy of text classification, our proposed method has certain advantages compared with the latest methods, but it requires further experimental verification to determine the reliability and stability of the algorithm. Therefore, we designed an experiment to investigate the stability of our proposed method using a small sample. We performed experiments on training sets of different scales to test the stability and reliability of the proposed model and benchmark algorithm.

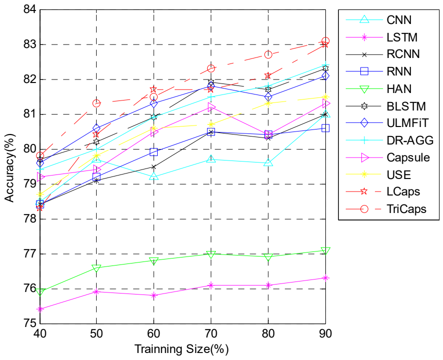

Fig. 4. Performance of different Text Classification methods with different training size using MR

We selected data x% as the training set, and the remaining data $(1 - x)\%$ as the test set. The x% value range is

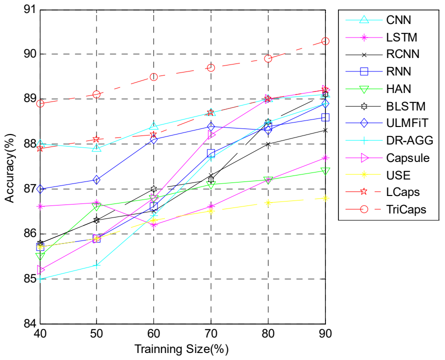

Fig. 5. Performance of different Text Classification methods with different training size using QA

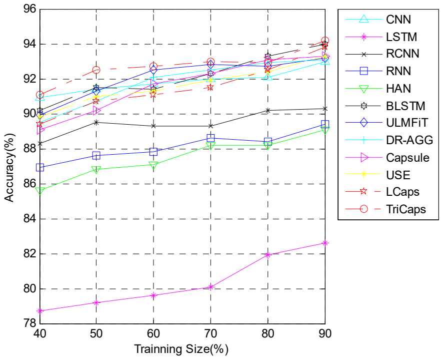

Fig. 6. Performance of different Text Classification methods with different training size using SU

$(40, 50, 60, 70, 80, 90)\%$. Accordingly, the value range of the test set was $(60, 50, 40, 30, 20, 10)\%$. The experimental results of the stability test for six datasets including MR,QA,SU and S1 are shown in Figure 4 to 7.

From Figure 4 to 7, it can be seen that our experimental results on four datasets show that our model is better than other benchmark algorithms for different training set sizes, and the overall average classification accuracy is higher than approximately 1% of the benchmark algorithm. The selection of the training set ranges from 40% to 90%, and as the size of the training set increases, our model shows steady growth. When the size of the training set is 90%, our model achieves the best performance. These results show that our proposed model is not only better than other benchmark algorithms, but also very stable for different training set sizes. These results shows that our proposed model is reliable and stable.

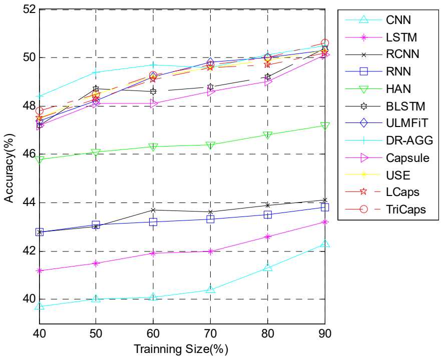

Fig. 7. Performance of different Text Classification methods with different training size using S1

V. CONCLUSION

In this paper, first, with the addition of the basic GRU unit, a basic capsule network is proposed to text classification, and the results are better than those of the CNN algorithm. Based on this, three capsule networks that share parameters are spatially combined to form a triplet network structure. Through two-stage training, global features of different samples are learned. The proposed model has excellent performance in the four datasets, indicating that global feature acquisition can improve the classification performance of the model to a certain extent. To the best of our knowledge, this is the first study to apply triplet loss to text classification. Through comparative learning, the neural network learned the global semantic differences between different categories. The previous methods only considered local feature information. In contrast, our method further improved the performance of the algorithm by comprehensively considering local and global features. The experimental results show that the proposed method considerably improves the classification accuracy of the model.

REFERENCES

[1] K. He and M. Zhu, "Text classification using gated and transposed attention networks," In Proceedings of IJCNN, 2019, pp.1-7.

[2] J. Cao, Z. Wu, J. Wu and H. Xiong, "SAIL: Summation-bAsed Incremental Learning for Information-Theoretic Text Clustering," IEEE Transactions on Cybernetics, 2013, 43(2): 570-584.

[3] J. Cao and Z. Wu, "An Improved Protocol for Deadlock and Livelock Avoidance Resource Co-allocation in Network Computing," World Wide Web Journal: Internet and Web Information Systems, 2010, 13(3): 373-388.

[4] L. Gao, J. Wu, C. Zhou and Y. Hu, "Collaborative Dynamic Sparse Topic Regression with User Profile Evolution for Item Recommendation," In Proceedings of AAAI,2017, pp.1316-1322.

[5] W. Yih, X. He and C. Meek, "Semantic parsing for single-relation question answering," In Proceedings of ACL, 2014, pp.643-648.

[6] Y. Shen, X. He, J. Gao, L. Deng and G. Mesnil. "Learning semantic representations using convolutional neural networks for web search," In Proceedings of WWW ,2014, pp.373-374.

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

[8] P. Liu, X. Qiu and X. Huang, "Recurrent neural network for text classification with multi-task learning," Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16): Human Language Technologies, 2017, pp.1480-1489.

[9] S. Sabour, N. Frosst and G. E. Hinton,"Dynamic routing between capsules," in Proc. Adv. Neural Inf. Process. Syst.,2017, pp.3859-3869.

[10] J. Gong, X. Qiu, S. Wang and X. Huang, "Information aggregation via dynamic routing for sequence encoding," In Proceedings of COLING, 2018, pp.2742-2752.

[11] M. Yang, W. Zhao, J. Ye, Z. Lei, Z. Zhao and S. Zhang,"Investigating Capsule Networks with Dynamic Routing for Text Classification," In Proceedings of EMNLP 2018, pp.3110-3119.

[12] M. Yang, W. Zhao, L. Chen, Q. Qu, Z. Zhao and Y. Shen, “Investigating the transferring capability of capsule networks for text classification,” Neural networks: the official journal of the International Neural Network Society, vol. 118, pp. 247-261, 2019.

[13] W. Zhao, H. Peng, S. Eger, E. Cambria and M. Yang, “Towards Scalable and Reliable Capsule Networks for Challenging NLP Applications,” In Proceedings of ACL, 2019, pp.1549-1559.

[14] Y. Wang, A. Sun, J. Han, Y. Liu and X. Zhu,“Sentiment analysis by capsules,” In Proceedings of WWW, 2018, pp.1165-1174.

[15] Z. Chen and T. Qian, “Transfer capsule network for aspect level sentiment classification,” In Proceedings of ACL, 2019, pp.547-556.

[16] T. Mullen and N. Collier,“Sentiment analysis using support vector machines with diverse information sources,” In Proceedings of EMNLP, 2004, pp.412-418..

[17] S. Tan, X. Cheng, Y. Wang and H. Xu,“Adapting naive Bayes to domain adaptation for sentiment analysis,” In Proceedings of ECIR, 2009, pp.337-349.

[18] J. Wu, Z. Cai, S. Zeng and X. Zhu,“Artificial immune system for attribute weighted Naive Bayes classification,” In Proceedings of IJCNN, 2013, pp.1-8.

[19] B. Trstenjak, S. Mikac and D. Donko, “KNN with TF-IDF based framework for text categorization,” Procedia Engineering, vol. 69, no. 1, pp. 1356C1364, 2014.

[20] J. Wu, S. Pan, X. Zhu, C. Zhang and X. Wu, “Multi-Instance Learning with Discriminative Bag Mapping,” IEEE Trans. Knowl. Data Eng, vol.30, no.1, pp.1065-1080, 2018.

[21] W. Lu, C. Zhou and J. Wu, “Big social network influence maximization via recursively estimating influence spread,” Knowl. Based Syst. vol.113, pp.143-154,2016.

[22] S. V. Wawre and S. N. Deshmukh,“Sentiment classification using machine learning techniques,” Int. J. Sci. Res., vol. 5, no. 4, pp. 819-821, 2016.

[23] R. Johnson and T. Zhang. “Effective use of word order for text categorization with convolutional neural networks,” In Proceedings of HLT-NAACL,2015, pp.103-112.

[24] D. Tang, B. Qin and T. Liu. “Learning semantic representations of users and products for document level sentiment classification,” In Proceedings of ACL, 2015, pp.1014-1023.

[25] A. M. Dai and Q. V. Le, “Semisupervised sequence learning,” In Proceedings of NurIPS, 2015, pp.3079-3087.

[26] S. Lai, L. Xu, K. Liu and J. Zhao, “Recurrent convolutional neural networks for text classification,” In Proceedings of AAAI,2015, pp.2267-2273.

[27] T. Miyato, A. M. Dai and I. Goodfellow, “Adversarial training methods for semi-supervised text classification,” In Proceedings of the International Conference on Learning Representations (ICLR), 2017, pp.1-11.

[28] B. Huang, Y. Ou and K. M. Carley. “Aspect Level Sentiment Classification with Attention-over-Attention Neural Networks,” In Proceedings of SBP-BRiMS, 2018, pp.197-206.

[29] Z Yang, D Yang, C Dyer, X He, A J. Smola and E H. Hovy, “Hierarchical Attention Networks for Document Classification,” In Proceedings of HLT-NAACL, 2016, pp.1480-1489.

[30] T. Shen, T. Zhou, G. Long, J. Jiang and C. Zhang, “Bi-directional block self-attention for fast and memory-efficient sequence modeling,” In Proceedings of ICLR, 2018, pp.1-18.

[31] Y. Lin, S. Shen, Z. Liu, H. Luan and M. Sun, “Neural relation extraction with selective attention over instances,” In Proceedings of ACL,2016, pp.2124-2133.

[32] B. Wang, “Disconnected recurrent neural networks for text categorization,"In Proceedings of ACL, 2018, pp.2311-2320.

[33] J. Howard and S. Ruder, “Universal language model fine-tuning for text classification,"In Proceedings of ACL, 2018, pp.328-339.

[34] J. Zeng, J. Li, Y. Song, C. Gao, M. R. Lyu and I. King, “Topic Memory Networks for Short Text Classification,” In Proceedings of EMNLP, 2018, pp.3120-3131.

[35] R. Johnson and T. Zhang, “Deep pyramid convolutional neural networks for text categorization,” In Proceedings of ACL, 2017, pp.562-570.

[36] X. Zhang, J. Zhao and Y. LeCun, “Character-level convolutional networks for text classification,” In Proceedings of NurIPS, 2015, pp.649-657.

[37] N. Kalchbrenner, E. Grefenstette and P. Blunsom, “A convolutional neural network for modelling sentences,” In Proceedings of ACL, 2014, pp.655-665.

[38] Q. Le and T. Mikolov, “Distributed representations of sentences and documents,” In Proceedings of ICML, 2014, pp.1188-1196.

[39] T. Mikolov, I. Sutskever, K. Chen, G. S Corrado and J. Dean, “Distributed representations of words and phrases and their compositionality,” In Proceedings of NurIPS, 2013, pp.3111-3119.

[40] J. Pennington, R. Socher and C. Manning, “Glove: Global vectors for word representation,” In Proceedings of EMNLP, 2014, pp.1532-1543.

[41] G. Wang, C. Li, W. Wang, Y. Zhang, D. Shen, X. Zhang, R. Henao and L. Carin. “Joint embedding of words and labels for text classification,” In Proceedings of ACL, 2018, pp.2321-2331.

[42] G. E. Hinton, A. Krizhevsky and S. D. Wang,“Transforming autoencoders,” in Proc. Int. Conf. Artif. Neural Netw. Springer, 2011, pp. 44-51.

[43] G. E. Hinton, S. Sabour and N. Frosst,“Matrix capsules with EM routing,” In Proceedings of ICLR, 2018, pp.1-12.

[44] E. Xi, S. Bing and Y. Jin,“Capsule network performance on complex data,” CoRR abs/1712.03480, 2017.

[45] A.Jaiswal, W. AbdAlmageed, Y. Wu and P. Natarajan, “CapsuleGAN: Generative adversarial capsule network,” In Proceedings of ECCV Workshops,2018, pp.526-535.

[46] S. Verma and Z. L. Zhang, “Graph capsule convolutional neural networks,” CoRR abs/1805.08090,2018.

[47] R. Lalonde and U. Bagci, “Capsules for object segmentation,” CoRR abs/1804.04241, 2018.

[48] B. Zhang, X. Xu, M. Yang, X. Chen and Y. Ye, “Cross-domain sentiment classification by capsule network with semantic rules,” IEEE Access,vol.6, pp.58284-58294,2018.

[49] Liqiang Xiao, Honglun Zhang, Wenqing Chen, Yongkun Wang, Yaohui Jin, “MCapsNet: Capsule Network for Text with Multi-Task Learning,” EMNLP 2018: 4565-4574.

[50] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Ion Androutsopoulos, “Large-Scale Multi-Label Text Classification on EU Legislation,” ACL 2019: 6314-6322.

[51] Hui Liu, Qingyu Yin, William Yang Wang, “Towards Explainable NLP: A Generative Explanation Framework for Text Classification,” ACL 2019: 5570-5581.

[52] Suchin Gururangan, Tam Dang, Dallas Card, Noah A. Smith, “Variational Pretraining for Semi-supervised Text Classification,” ACL 2019: 5880-5894.

[53] E. Hoffer and N. Ailon, “Deep metric learning using triplet network,” In Proceedings of SIMBAD, 2015, pp.84-92.

[54] B. Pang and L. Lee, “Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales,” In Proceedings of ACL, 2005, pp.115-124.

[55] J. Wiebe, T. Wilson and C. Cardie, “Annotating expressions of opinions and emotions in language,” Language Resources and Evaluation, vol.39, pp.165-210, 2005.

[56] B. Pang and L. Lee, “A sentimental education: Sentiment analysis using Subjectivity datasets ectivity summarization based on minimum cuts,” In Proceedings of ACL,2004, pp.271-278.

[57] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng and C. Potts, “Recursive deep models for semantic compositionality over a sentiment treebank,” In Proceedings of EMNLP,2013, pp.1631-1642.

[58] X. Li and D. Roth, “Learning question classiers,” In Proceedings of COLING, 2002.

[59] K. Cho, B. v. Merrenboer, C. G. D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio,“Learning phrase representations using rnn encoder-decoder for statistical machine,” In Proceedings EMNLP, 2014, pp. 1724-1734.

[60] P. Zhou, Z. Qi, S. Zheng, J. Xu, H. Bao and B. Xu, “Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling,” In Proceedings of COLING, 2016, pp. 3485-3495.

[61] D. Cer, Y. Yang, S. yi Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, B. Strope and R. Kurzweil, “Universal sentence encoder for english,” In Proceedings of EMNLP, 2018, pp. 169-174.