Gated iterative capsule network for adverse drug reaction detection from social media

Tongxuan Zhang[1], Hongfei Lin[1*], Bo Xu[1,2], Yuqi Ren[1], Zhihao Yang[1], Jian Wang[1], Xiaodong Duan[3]
[1] *College of Computer Science and Technology, Dalian University of Technology*
[2] *State Key Laboratory of Cognitive Intelligence,iFLYTEK, P.R.China*
[3] *College of Computer Science and Engineering, Dalian Minzu University*
Dalian, China
hflin@dlut.edu.cn

2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 978-1-7281-6215-7/20/$31.00 ©2020 IEEE DOI: 10.1109/BIBM49941.2020.9313092

Abstract—**In this paper, we propose a gated iterative capsule network model for the ADR detection task, named GICN. To alleviate the impact caused by abbreviations and misspelled words, we add character embedding as part of the input. Most ADRs consist of multiple words, e.g., short-term memory dysfunction. Hence, we apply a convolutional neural network (CNN) to obtain the complete phrase information. To effectively extract deep semantic information, we introduce a capsule network with a gated iteration unit that clusters features from underlying to high capsules. The gated iteration mechanism can remember contextual information, which will be introduced when clustering features. Experimental results show that our approach can achieve significant performance improvement for ADR detection from social media text compared with other state-of-the-art works.**

Keywords—adverse drug reactions; capsule network; gated iteration unit; social media

I. Introduction

With the increase in drugs worldwide, it is difficult to completely determine the effects of drugs on clinical response. Adverse drug reactions (ADRs) refer to the dangerous effects that may occur during drug use. Drugs usually go through clinical trials before being approved for use. However, it is impossible to uncover all ADRs during trials, which will take a long period and a number of experiments. Many researchers have extracted potential ADRs from text that records the state of patients after medication, such as medical literature [1-2] and social media [3-5] about ADRs. Therefore, natural language processing (NLP) technology [5-8] is very important in automatically detecting ADRs from text.

Compared with the medical literature, ADRs on posted on social media in a timely manner by patients are more valuable for researching unknown ADRs. Social media provides a platform for consumers to share their experiences and opinions about feelings after taking medicine. Hence, a large number of researchers use social media texts as a data source for ADR detection [3-5]. However, due to the freedom of social media, people freely express their opinions on social media, regardless of whether their sentences are complete sentences or correct grammar is used.

For example, the language of social media is informal. It contains a lot of colloquial descriptions (eg, “my sleep was worse than normal”, “hands still shaking”). It also includes misspelling (eg, “alwwpong”, “havee”). Netizens do not have a comprehensive understanding of medical terms. They use abbreviations to convey semantic information (eg. “sp?”, “ADHD”). However, these problems should not be the reasons that hinder the use of social media data to carry out ADR detection research.

Generally, there are two main challenges to ADR detection from social media: 1) social media texts include irregular grammar and various expressions, and 2) there are some misspelled words in sentences. To solve the above problems, we improve on original capsule networks [9] and propose a gated iterative capsule network model (GICN). Compared with most existing neural network models for ADRs, sentence representations can contain richer semantic and structural features by GICN. GICN can also learn memory information in a sentence sequence by a gated iteration unit, which reserves important contextual information for ADR detection. We also use a character CNN (charCNN) to address the spelling mistakes. In this paper, we focus on text extracted from social media.

II. Methods

The motivation for this research is whether we can design a capsule network to learn deep semantic information. First, the embedding layer consists of three parts: character embedding, word embedding and position embedding. Then, we use CNN to capture phrase features. Finally, a gated iterative capsule network is applied to learn high-level semantic representation considering the parts of the drug and symptoms with memory information, Figure 1 illustrates our approach applied to the relation extraction of ADRs.

A. Embedding input representation

We take a sentence sequence S, $\{w_1, w_2 ..., w_n\}$ as input. Each word w_n is represented by a vector $E_{w_i} = [E_{word} + E_{c\hbar ar} + E_{pos1} + E_{pos2}]$. The vector E_{w_i} has four parts: the word embedding $E_{word} \in R^{d_{word}}$, the character embedding $E_{char} \in R^{d_{ch}}$ and the position embedding $E_{pos} \in R^{d_p}$. ch denotes the dimension of length w_n and p denotes the dimension of position embedding. The vector is $E_{w_i} \in R^{d_{word+ch+p1+p2}}$.

B. Feature extracting layer

We use a convolutional neural network (CNN) [10] as a feature extractor to capture phrase features and perform sentence representations. The formula of CNN function is:

$$u_t = f(W_C \cdot X_{t:t+\hbar-1} + b) \tag{1}$$

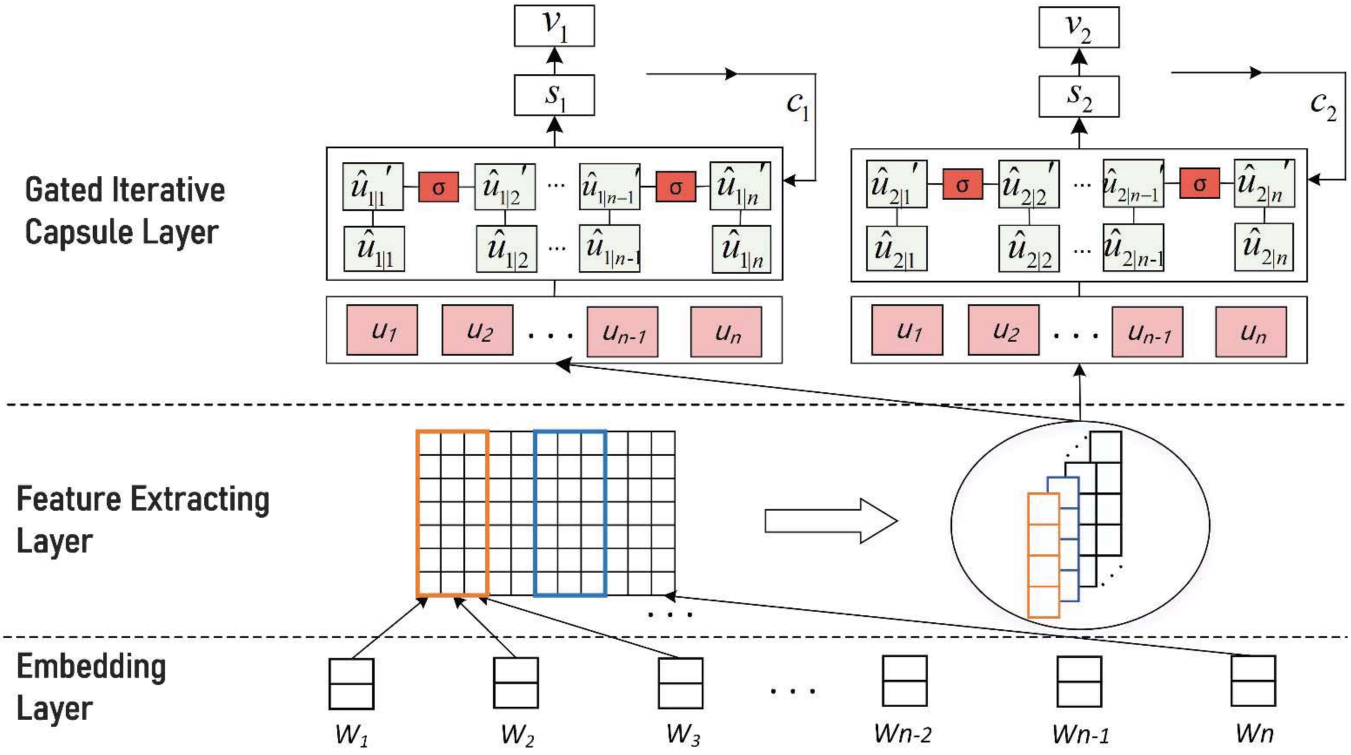

Figure 1. The architecture of our proposed gated iterative capsule network model

where f denotes an activation function. W_C is a weight matric and b is a bias vector. We use a fixed size window $\hbar$ to integrate phrase information. $X_{t:t+\hbar-1}$ denotes the word x from *t-th* to *(t+h-1)-th* from input token E_{w_t}. The new matrix $U = \left(u_1, u_2 \ldots, u_{n-\hbar+1}\right)$ reflects the more expressive local semantic meaning of the sentence. $U \in R^{n' \times d}$ where $n' = n - \hbar + 1$ is sequence length, h is the window size and d is the dimensional size of CNN.

C. Gated iterative capsule network

Hinton et al. [11] originally proposed the capsule network in the field of visual images for digit recognition. In our work, we propose a gated iterative capsule network with a gated iteration unit for the ADR detection task. Between the two layers, the layer *l* capsule output u_{i} produce the new vector $\hat{u}_{j|i}$.

$$\hat{u}_{j|i} = W_{ij} u_{\mathrm{i}} \tag{2}$$

where W_{ij} indicates a weight matrix.

However, different words in the sentence have different contributions to classification. Inspired by a gate mechanism [12], we propose the gated iterative capsule network. The second layer vectors $\hat{u}_{j|i}{}'$ is produced by sharing the prior feature and obtain the memory information among the $\hat{u}_{j|i}$. A gate g is calculated from the prior $\hat{u}_{j|i-1}{}'$ as follow:

$$g_{j|i} = \sigma(W_g \hat{u}_{j|i-1}{}' + b_g) \tag{3}$$

where σ denotes the nonlinear activating of sigmoid. W_g indicates a weight matrix and b_g indicates a bias vector. The vector gate g can select the helpful information. Therefore, the $\hat{u}_{j|i}{}'$ is calculated by the lower layers $\hat{u}_{\mathrm{j|i}}$ and prior $\hat{u}_{j|i-1}{}'$.

$$\hat{u}_{j|i}{}' = g_{j|i} \odot \hat{u}_{j|i-1}{}' + \hat{u}_{\mathrm{j|i}} \tag{4}$$

where $\odot$ is element-wise multi-plication. $\hat{u}_{j|i}$ produced by the capsule u_{i} in layer *l*. In the calculation, the $\hat{u}_{j|1}{}'$ is $\hat{u}_{\mathrm{j|1}}$. In our model, the $\hat{u}_{j|i}{}'$ is including the memory information. Then, the input of the capsule s_j in layer *l+1* is produced by all $\hat{u}_{j|i}{}'$ through the weighted sum. The memory information in the sequence is obtained by s_j:

$$s_j = \sum_i c_{ij} \, \hat{u}_{j|i}{}' \tag{5}$$

where the c_{ij} indicates the coupling coefficient. The iterative dynamic routing algorithm determines c_{ij} . Furthermore, the output v_j of layer *l+1* is obtained by the following formula which is a non-linear squashing function:

$$v_j = \frac{\|s_j\|^2}{1+\|s_j\|^2} \frac{s_j}{\|s_j\|} \tag{6}$$

where v_j is the output of *l+1* capsule. v_j value represents the probability of each category. Thus, by the non-linear squashing function, the output v_j is limited in range [0, 1]. The short vectors are shrunk to zero length, and the long vectors are shrunk to one length.

D. Dynamic routing algorithm

Through the dynamic routing algorithm, we can obtain the high-level vector representation which represents relation features. The algorithm is described in Algorithm 1.

Algorithm 1: Dynamic Routing Algorithm

```
procedure ROUTING($\hat{u}_{j|i}'$, $r$, $l$)
for the capsule $i$ in layer $l$ and the capsule $j$ in layer $l+1$:
  initialize the logits of coupling coefficients $b_{ij} = 0$
for $r$ iterations do
    $c_i = softmax(b_j)$
    $s_j = \sum_i c_{ij} \hat{u}_{j|i}'$
    $v_j = g(s_j)$   non-linear squashing function
    $b_{ij} = b_{ij} + \hat{u}_{j|i}' \cdot v_j$
end for
return $v_j$
```

Where r is the number of iterations. The coupling coefficient c between all the layer l capsule i and all the layer $l+1$ capsule j. By the softmax function, initial logits b_j determine the coupling coefficient c.

E. Margin loss

In this paper, we apply the margin loss for the classification ADRs as Sabour et al. [13]. For each relation category capsule v_j, the margin loss is as follow:

$$L_j = Y_j max\left(0, m^+ - \|v_j\|\right)^2 \\ +\lambda\left(1 - Y_j\right) max(0, \|v_j\| - m^-)^2 \quad (7)$$

If the sentence represents the corresponding relation, $Y_j = 1$. And if the sentence doesn't represent the corresponding relation, $Y_j = 0$. λ indicates the weight for the absent classes, m^+ and m^- are the top and bottom margin. In our experiment, $\lambda = 0.5$, $m^+ = 0.9$ and $m^- = 0.1$.

III. Experiment

A. Experimental datasets and settings

We test our method on TwiMed [14]. The summary statistics are presented in Table 1.

TABLE I. SUMMARY STATISTICS OF CORPUS

Coupus	Documents	ADR	non-ADR	Max length	Experimental length
TwiMed-Twitter	625	311	301	64	60
TwiMed-PubMed	1000	264	983	137	70

We implemented our method by Keras. The dimension of the word embedding was 100-dimensional by pretrained. The position embedding was 10-dimensional by random initialization. The dimension of the character embedding was 25-dimensional. To optimize the parameters, we applied the Adam optimizer, and the learning rate was 0.01. The number of CNN layer hidden units was 64, and recurrent dropout was 0.5. The maximum epoch was 30, and the batch size was 8. The capsule dimension was 32. The iteration r was 3. We used precision (P), recall (R) and F-score (F1) as evaluation metrics for ADR detection in our experiment. We evaluated all models by 10-fold cross-validation.

B. Evaluation of TwiMed

In Table 2, Alimova et al. [16] proposed the first two line models, which are the main models and baselines. The first-line method is a combination of feature-rich SVM and linear kernel [15]. The interactive attention network (IAN) [16] learns the contextual feature. Zhang et al. [3] proposed a multi-hop self-attention mechanism (MSAM) model that learns the multi-aspect semantic information for ADR detection. We also implemented some methods in the next three lines of Table 2. Some researchers [17-18] used CNN-based methods with max-pooling on the drug-drug interaction (DDI) extraction task. Kumar et al. [1] used an LSTM-based joint AB-LSTM model with max-pooling for the DDI task. By max-pooling, using a fixed-size vector to represent sentence information lacks guidance using the task. This is also the main factor associated with unsatisfactory results.

The experimental results of the last two lines are the results of our proposed GICN model. The GICN-charCNN method represents the GICN model without character embeddings, for which performance degradation is obvious. Experimental results show that character embeddings can learn character features, which is beneficial to classification. The performance improvement on TwiMed-Twitter is much better than the performance improvement on TwiMed-PubMed. Due to the nonstandard language of Twitter, there are many language spelling problems, so our model with character embedding contributes to performance improvement. The GICN in the last line obtains memory information through the gated iteration unit in the capsule and performs a high-dimensional feature representation of the sentence. It integrates the phrase feature and memory feature as the final classification basis. Compared with other baseline models in ADR detection, our result shows that the GICN performs better. In terms of our model results, neither precision nor recall is optimal. However, we improved both of them. We narrowed the gap between precision and recall, thereby increasing the F1 score. Compared with the MSAM [3], our model gained a 0.8% improvement in the F1 score on TwiMed-PubMed. Compared with the IAN [16], our model gained a 1.3% improvement in the F1 score on TwiMed-Twitter.

IV. Conclusion

In this paper, we propose a gated iterative capsule network (GICN) for ADR detection. The main idea of the proposed model is extracting high-level semantic information to focus on ADR phrases without requiring any linguistic knowledge. By the internal gating mechanism, our model can learn memory information from a sentence sequence. Moreover, char CNN can obtain information between the characters, which is helpful for ADR detection. Experiments on a real-world dataset show the effectiveness of the proposed models when compared with other alternatives as well as existing methods.

TABLE II. CLASSIFICATION RESULTS OF THE COMPARED METHODS FOR TWIMED CORPUS.

Method	TwiMed-PubMed			TwiMed-Twitter		
	P	R	F1	P	R	F1
Feature-rich SVM [15]	0.799	0.681	0.728	0.752	0.810	0.778
IAN[16]	0.878	0.738	0.792	**0.836**	0.813	0.824
MSAM [3]	0.858	0.852	0.853	0.748	**0.856**	0.799
CNN-based method [17]#	0.849	0.831	0.835	0.739	0.788	0.761
multichannel CNN [18] #	0.861	0.780	0.816	0.738	0.841	0.780
Joint AB-LSTM [1] #	0.817	**0.856**	0.831	0.701	0.828	0.754
GICN-charCNN	0.864	0.835	0.847	0.826	0.797	0.809
GICN	**0.879**	0.852	**0.861**	0.834	0.841	**0.837**

Models with # are our implementations.

REFERENCES

[1] Sahu S K, Anand A. Drug-drug interaction extraction from biomedical texts using long short-term memory network[J]. Journal of biomedical informatics, 2018, 86: 15-24.

[2] Wei C H, Peng Y, Leaman R, et al. Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task[J]. Database, 2016, 2016.

[3] Zhang T, Lin H, Ren Y, et al. Adverse drug reaction detection via a multihop self-attention mechanism[J]. BMC bioinformatics, 2019, 20(1): 479.

[4] Li Z, Yang Z, Luo L, et al. Exploiting Adversarial Transfer Learning for Adverse Drug Reaction Detection from Texts[J]. Journal of Biomedical Informatics, 2020: 103431.

[5] Lee K, Qadir A, Hasan S A, et al. Adverse drug event detection in tweets with semi-supervised convolutional neural networks[C]//Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2017: 705-714

[6] K. Lee, A. Agrawal, and A. Choudhary. Mining social media streams to improve public health allergy surveillance. In 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pages 815–822, Aug 2015.

[7] Nikfarjam A, Sarker A, O'Connor K, et al. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features[J]. Journal of the American Medical Informatics Association, 2015, 22(3): 671-681.

[8] Xu J, Wu Y, Zhang Y, et al. CD-REST: a system for extracting chemical-induced disease relation in literature[J]. Database, 2016, 2016.

[9] Sabour S, Frosst N, Hinton G E. Dynamic routing between capsules[C]//Advances in neural information processing systems. 2017: 3856-3866.

[10] Zeng D, Liu K, Lai S, et al. Relation classification via convolutional deep neural network[C]//Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. 2014: 2335-2344.

[11] Hinton G E, Krizhevsky A, Wang S D. Transforming auto-encoders[C]//International Conference on Artificial Neural Networks. Springer, Berlin, Heidelberg, 2011: 44-51.

[12] Xiao L, Zhang H, Chen W. Gated Multi-Task Network for Text Classification[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). 2018: 726-731.

[13] Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records[J]. NPJ Digital Medicine, 2018, 1(1): 18.

[14] Alvaro N, Miyao Y, Collier N. TwiMed: Twitter and PubMed comparable corpus of drugs, diseases, symptoms, and their relations[J]. JMIR public health and surveillance, 2017, 3(2): e24.

[15] Alimova I, Tutubalina E. Automated detection of adverse drug reactions from social media posts with machine learning[C]//International Conference on Analysis of Images, Social Networks and Texts. Springer, Cham, 2017: 3-15.

[16] Alimova I, Solovyev V. Interactive Attention Network for Adverse Drug Reaction Classification[C]//Conference on Artificial Intelligence and Natural Language. Springer, Cham, 2018: 185-196.

[17] Liu S, Tang B, Chen Q, et al. Drug-drug interaction extraction via convolutional neural networks[J]. Computational and mathematical methods in medicine, 2016, 2016.

[18] Quan C, Hua L, Sun X, et al. Multichannel convolutional neural network for biological relation extraction[J]. BioMed research international, 2016, 2016.