BDCN: Semantic Embedding Self-explanatory Breast Diagnostic Capsules Network

Dehua Chen, keting Zhong, Jianrong He*
School of Computer Science and Engineering, Donghua University, Shanghai, China
Comprehensive Breast Health Center, Shanghai Ruijin Hospital, Shanghai, China
chendehua@dhu.edu.cn {katyzhong7,hejrongbreastsg}@163.com

Abstract

Building an interpretable AI diagnosis system for breast cancer is an important embodiment of AI assisted medicine. Traditional breast cancer diagnosis methods based on machine learning are easy to explain, but the accuracy is very low. Deep neural network greatly improves the accuracy of diagnosis, but the black box model does not provide transparency and interpretation. In this work, we propose a semantic embedding self-explanatory Breast Diagnostic Capsules Network (BDCN). This model is the first to combine the capsule network with semantic embedding for the AI diagnosis of breast tumors, using capsules to simulate semantics. We pre-trained the extraction word vector by embedding the semantic tree into the BERT and used the capsule network to improve the semantic representation of multiple heads of attention to construct the extraction feature, the capsule network was extended from the computer vision classification task to the text classification task. Simultaneously, both the back propagation principle and dynamic routing algorithm are used to realize the local interpretability of the diagnostic model. The experimental results show that this breast diagnosis model improves the model performance and has good interpretability, which is more suitable for clinical situations.

1 Introduction

Breast cancer is an important killer threatening women's health because of rising incidence. Early detection and diagnosis are the key to reduce the mortality rate of breast cancer and improve the quality of life of patients. Mammary gland molybdenum target report contains rich semantic information, which can directly reflect the results of breast cancer screening (CACA-CBCS, 2019), and AI-assisted diagnosis of breast cancer is an important means. Therefore, various diagnostic models were born. Mengwan (2020) used support vector machine(SVM) and Naive Bayes to classify morphological features with an accuracy of 91.11%. Wei (2009) proposed a classification method of breast cancer based on SVM, and the accuracy of the classifier experiment is 79.25%. These traditional AI diagnoses of breast tumors have limited data volume and low accuracy. Deep Neural Networks (DNN) enters into the ranks of the diagnosis of breast tumor. Wang (2019) put forward a kind of based on feature fusion with CNN deep features of breast computer-aided diagnosis methods, the accuracy is 92.3%. Zhao (2018) investigated capsule networks with dynamic routing for text classification, which proves the feasibility of text categorization. Existing models have poor predictive effect and lack of interpretation, which can not meet the clinical needs.

Based on the above pain points, we propose a semantic embedding self-explanatory Breast Diagnostic Capsules Network (BDCN), which diagnoses breast tumors based on the mammary gland molybdenum target report. Our contributions are as follows:

- Semantic segmentation algorithm is used to segment breast cancer lesions.
- Semantic tree is integrated into Bidirectional Encoder Representation from Transformers(BERT) pre-training to obtain word vectors.

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1178-1189, Hohhot, China, Augest 13 - 15, 2021.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

- A capsule network with multi-head attention mechanism was proposed to predict breast tumors.
- Using Back Propagation to Realize Local Interpretation of BDCN Model.

2 Related Works

A Self-interpretation Method based on Capsule Network: LIME (Ribeiro et al., 2016) realizes partial interpretability of the model based on the idea of perturbation. SHAP (Lundberg and Lee, 2017) treats all features as ”contributors” and produces a predicted value. These can be explained in hindsight, they were a unified interpretation method for all models. We do not consider the model independent interpretation method, through back propagation and weight sharing (Wang et al., 2020) to construct a semantic embedding self-explanatory Breast Diagnostic Capsules Network. Capsule network (Sabour, 2017) with the dynamic routing algorithm to dynamically determine rights, itself from a certain extent. It has provided the edge explanatory power to determine, which is an interpretable model based on network node association analysis.

Fusion Embedded for Semantic Vector: word2vec (Mikolov et al., 2013) is a representative of the embedded word paradigm, but the word that produces word2vec is static, regardless of the relationship between context and connection. BERT (Devlin et al., 2018) provides many pre-training models, which are excellent in various evaluation indexes. However, since BERT is trained in open data sets, the word vectors directly extracted by Bert are not accurate for specific fields.

3 Model

BDCN is based on semantic embedding, multi-head attention and capsule network to realize text diagnosis of breast cancer examination report. The overall architecture is shown in Figure 1. It is mainly composed of three modules: the semantic segmentation layer of inspection report, the semantic tree knowledge embedding extraction word vector layer (Sem-Bert), and the capsule network assisted target feature multi-head attention representation classification layer (Muti-Cap).

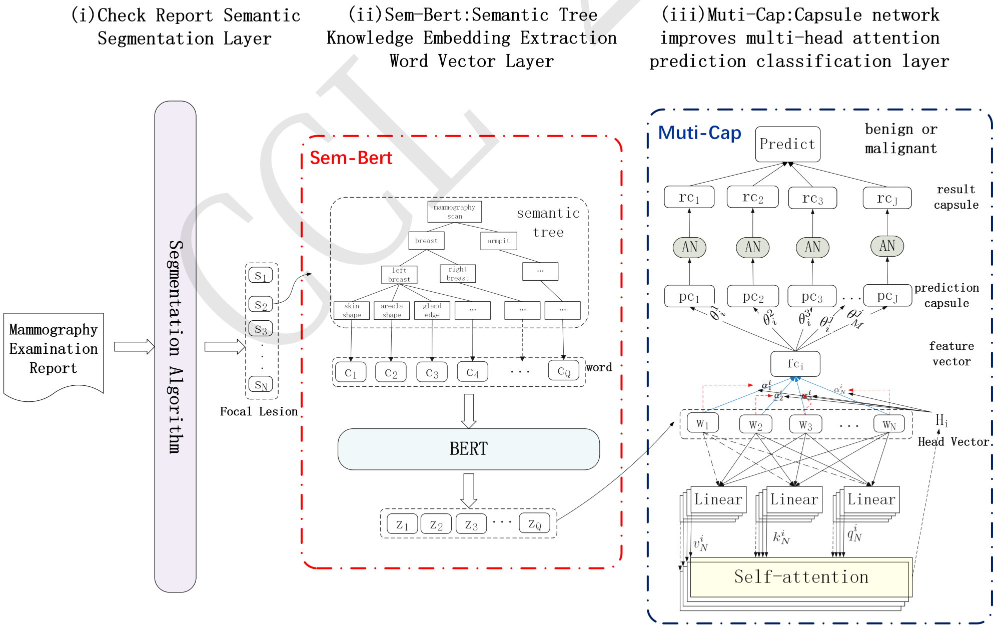

(i)Check Report Semantic Segmentation Layer
(ii)Sem-Bert:Semantic Tree Knowledge Embedding Extraction Word Vector Layer
(iii)Muti-Cap:Capsule network improves multi-head attention prediction classification layer
Mammography Examination Report
Segmentation Algorithm
S_1
S_2
S_3
S_N
Focal Lesion
Sem-Bert
mammography scan
semantic tree
breast
armpit
left breast
right breast
...
skin shape
areola shape
gland edge
c_1 c_2 c_3 c_4 ... c_Q
word
BERT
z_1 z_2 z_3 ... z_Q
Muti-Cap
Predict
benign or malignant
rc_1 rc_2 rc_3 rc_J
result capsule
AN
pc_1 pc_2 pc_3 ... pc_J
prediction capsule
θ_i^1 θ_i^2 θ_i^3 θ_i^j θ_M^j
fc_i
feature vector
α_1^i α_2^i α_3^i α_N^i
w_1 w_2 w_3 ... w_N
H_i
Head Vector
Linear
v_N^i k_N^i q_N^i
Self-attention

Figure 1. Overall architecture of the BDCN.

The input is an original mammography report. We use segmentation algorithm to attain preprocessing and segmentation of lesions and the implementation details are shown in Section 3.1. Then we propose Sem-Bert method, which uses semantic tree to improve BERT and pre-training to obtain the word vector for the medical field. The details are discussed in Section 3.2. The word embedding z_1,z_2,$\cdots$,$z_N \in R^{d_z}$

is converted into the phrase embedding w_1,w_2,$\cdots$,$w_N \in R^{d_w}$ in one-dimensional convolution. Next, the feature capsule fc_1,fc_2,$\cdots$,$fc_M \in R^{d_f}$ is transformed into a prediction capsule pc_j, $j \in [1, 3]$ based on the dynamic routing algorithm, and the result capsule rc_j, $j \in [1, 3]$ is obtained by activating the network. Finally, the benign and malignant diagnosis of breast tumors was predicted, and the implementation details are shown in Section 3.3.

3.1 Check the Report Semantic Segmentation Layer

The main role of this layer is to achieve semantic segmentation of the report. The clinician can diagnose the breast tumor by analyzing the mammography report to judge the pathological condition of the patient's breast and its surrounding tissue. However, a single report may contain multiple focal lesions in the same location or even in different gland background tissues, which may lead to lower prediction accuracy if generalized. Therefore, segmentation of breast report is an important link. The core steps of the segmentation algorithm are as follows:

- Rough segmentation. Divide the report into sentences based on ”” or ””, then according to different keywords such as skin, mass and axilla, they were divided into corresponding glandular background, focal lesions and axilla.

- Further subdivision. Most focal lesions mixed in the background part of the gland contain the keywords ”nodules” or ”densified shadow”. The sentence of the background part of the gland is further subdivided to screen out the information of focal lesions and axillary parts. The sentences on the glandular background were further subdivided to screen for information on focal lesions and axillary lesions.

- Distinguish the left and right sides according to the the position described. According to ”, ” is divided into short sentences, when the ”left” or ”right” keyword appears, the short sentences are divided according to the position words;When ”double” appears, it is divided into left and right sides.

3.2 Sem-Bert Layer

This layer is mainly to extract word vectors with BERT combined with semantic tree (Chen et al., 2019; Jiang and He, 2020) and the process diagram of Sem-BERT to obtain word vectors is shown in Figure 2. The Sem-Bert method constructs a semantic tree first. Semantic tree has obvious advantages of context hierarchy, so the construction of semantic tree can help to solve the problem of unreasonable word segmentation and context incoherence. According to the rule of ”segment - organization description sentence - attribute description sentence”, relational extraction is carried out, and dependency syntax is used to construct Extensible Markup Language(XML) semantic tree. The concrete content of report semantic tree construction includes the following five parts.

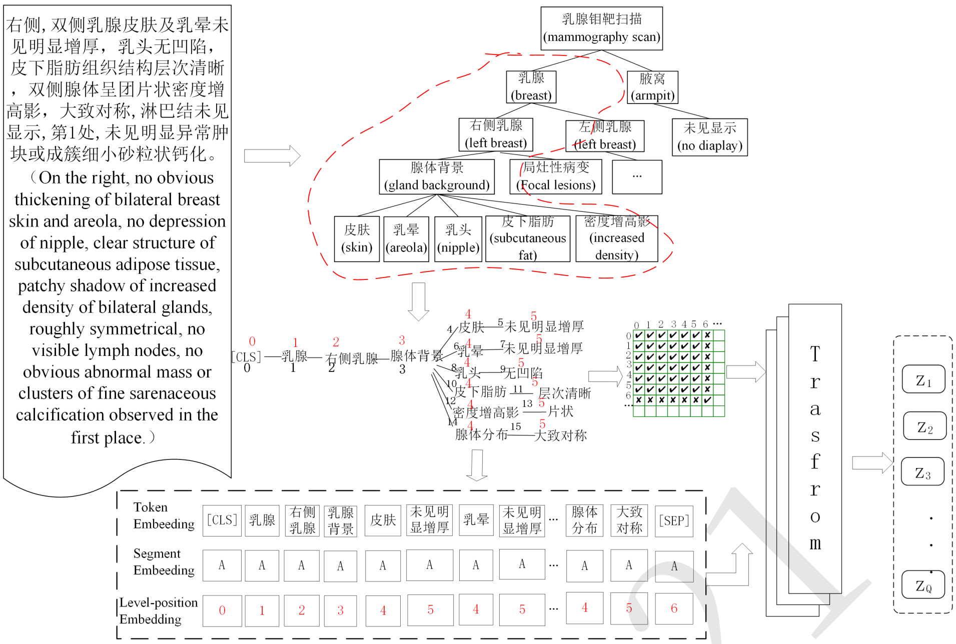
- Chinese word segmentation. Jieba word segmentation tool is the best choice to ensure the accuracy of word segmentation, and breast molybdenum target dictionary can be customized according to the knowledge of breast molybdenum target terminology combined with clinicians' guidance.

- Synonym conversion. Different doctors have different habits of describing mammograms. Replacing the words with the same general meaning in medical science into unified words can effectively reduce the redundancy of semantic tree.

- Organization description sentence acquisition. Find the organization description sentence of the corresponding part of the paragraph, scan the report from left to right, and the description before each organization word A is encountered until the next new organization word B is encountered will be classified as the description of A.

- Organization segment subtree path acquisition. Each organization descriptive sentence is converted into an attribute description sentence and the attribute value is extracted. Taking the current organization as the root node, we find finer attributes through dependency syntax and extract the attribute value of each attribute in the organization word.

- Mammary report dictionary construction and XML transformation. The information extracted from different parts was added to the branches of the semantic tree, and the duplicated information was pruned to obtain the mammary gland report dictionary.



右侧，双侧乳腺皮肤及乳晕未见明显增厚，乳头无凹陷，皮下脂肪组织结构层次清晰，双侧腺体呈团片状密度增高影，大致对称，淋巴结未见显示，第1处，未见明显异常肿块或成簇细小砂粒状钙化。

（On the right, no obvious thickening of bilateral breast skin and areola, no depression of nipple, clear structure of subcutaneous adipose tissue, patchy shadow of increased density of bilateral glands, roughly symmetrical, no visible lymph nodes, no obvious abnormal mass or clusters of fine sarenaceous calcification observed in the first place.）

Figure 2. The structure diagram of word vector was obtained by Sem-Bert method.

Taking the semantic tree as input, by configuring the BertConfig class, setting the Tokenizer word slicer, then numbering the words in the lexicon and converting them into dictionaries, selecting the word that makes the likelihood function increase the most, and selecting the word slicer according to the frequency. The structural diagram of the Sem-Bert method is shown in Figure 2, taking the gland background description information as an example (circled by a red dotted line). Similar to Bert, the mammography report semantic tree needs to be converted into a sequence by means of token embedding, position embedding and segment embedding, while preserving its structural information. However, unlike traditional BERT, because the input is a semantic tree rather than a sequence of tokens, the positional embedding of the BERT input needs to be changed in order to preserve the structural information of the breast examination report. Position bedding is changed to level-position embedding and original-position embedding, marked by red and black numbers in Figure 2, respectively. Level-position id represents the position of the same branch in the semantic tree, and gives each Token the same branch to scale the hierarchical order information of the semantic tree starting from 0. The Original-position id is represented in the same way as the position id in BERT. [SEP] represents a special marker for multiple sentences. However, in order to preserve the structural information, a relational matrix is introduced to record whether it is reachable under the same branch, reachable under the same branch, or not. Lastly, multiple self-attentions in Transform are stacked with each other to code, and the final word vector z_1,z_2,$\cdots$,$z_N \in R^{d_z}$ is obtained by pre-training.

3.3 Muti-Cap Layer

This layer converts the word vectors of the pre-training layer into capsules, uses the capsule network to obtain the required prediction capsules, and combines effective information from multiple attention heads to achieve better classification. As one of the three most powerful semantic feature extractors, transform's self - attention mechanism is superior to CNN and neural network in word sense ambiguity resolution (Long et al., 2015). However, when the vector dimension is too high, the self-attention in each component represents different features, which leads to the fact that all attention can not fully capture the features (Tang et al., 2018). Multi-head attention can learn the features of sequences from different aspects, which is helpful for the network to capture more abundant features. Each of these head vectors points to a feature capsule, performing attention independently. Long attention to extract the feature layer will get w_1,w_2,$\cdots$,$w_N \in R^{d_w}$ as its input in the layer. By way of generating M features capsule fc_1,fc_2,$\cdots$,$fc_M \in R^{d_f}$, we need to have M long since attention vector h_1,h_2,$\cdots$,$h_M \in R^{d_h}$, and for each Hi $H_i \in R^{d_f}$, $m \in [1, M]$ can obtain different key vectors k_n^i, value vectors v_n^i and query vectors q_n^i through linear transformation, and then get more attention by the vector series. The generation process of the head vector is as follows:

for n = 1, 2, 3, $\cdots$,N do

$$v_n^i = W_i^V w_n \tag{1}$$

$$k_n^i = W_i^K w_n \tag{2}$$

$$q_n^i = W_i^Q w_n \tag{3}$$

$$H_i = attention(q_n^i, k_n^i, v_n^i) \tag{4}$$

where $W_i^V \in R^{d_f \times d_w}$,$W_i^K \in R^{d_h \times d_w}$ are parameters. The weight of attention and the output feature capsule of multi-head attention can be calculated. The formulas are as following:

$$\alpha_n^i = \frac{exp(\frac{H_i^T k_n^i}{\sqrt{d_p}})}{\sum_{n'} exp(\frac{H_i^T k_{n'}^i}{\sqrt{d_p}})} \tag{5}$$

$$fc_i = \sum_n \alpha_n^i v_n^i \tag{6}$$

The feature capsule is obtained by the sum of attention weights. As a parameter, the number of feature capsules is adjustable, which solves the problem that the intermediate parameters of three prediction capsules and M feature capsules are too large, and we need to normalize the attention weight:

$$\sum_n \alpha_n^i = 1 \tag{7}$$

With CNN classification, some important information will be lost in the operation of the aggregation layer, while capsule network represents a group of neurons by capsule, replacing the neuron output vector (Sabour, 2017). Therefore, the capsule network can effectively represent the location and semantics of features, and each upper capsule is the high-level semantics of the lower capsule. In addition, it can improve the information aggregation of multiple attention, so as to obtain more effective features and improve the ability of text representation. The structural schematic diagram is shown in Figure 3.

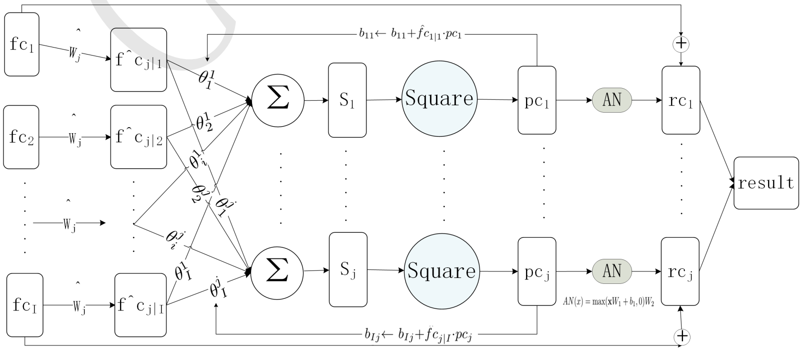

fc₁
fc₂
fc_I
Ŵ_j
f̂c_j|1
f̂c_j|2
f̂c_j|I
θ₁¹
θ₂¹
θ_i¹
θ₂ʲ
θ₁ʲ
θ_iʲ
θ_I¹
θ_Iʲ
$b_{11} \leftarrow b_{11} + \hat{f}c_{1|1} \cdot pc_1$
$b_{Ij} \leftarrow b_{Ij} + \hat{f}c_{j|I} \cdot pc_j$
Σ
S₁
S_j
Square
pc₁
pc_j
AN
rc₁
rc_j
+
result
$AN(x) = \max(xW_1 + b_1, 0)W_2$

Figure 3. Structure diagram of improved capsule network.

The dynamic routing algorithm of the capsule network (Zhao et al., 2019) is used to calculate the prediction capsule pc_i. Through multiple iterations of the routing process, Muti-Cap can determine the number of characteristic capsules flowing into the prediction capsules, which plays an significant role in the prediction capsules. Firstly, the feature capsule vector is calculated by inputting the feature capsule

pc_i and multiplying the learning transformation matrix $\hat{W}_j$. The dynamic routing weight θ_i^j is then determined by calculating the "routing softmax" of the initial logits b_{ij}, the formulas are as following:

$$\hat{fc}_{j|i} = \hat{W}_j fc_i + \hat{b}_{ij} \tag{8}$$

$$\theta_i^j = \frac{exp(b_{ij})}{\sum\limits_j exp(b_{ij})} \tag{9}$$

where $\hat{W}_j$ here is shared among each feature capsule to obtain the feature capsule vector, and b_{ij} is initialized to 0.

The final output is normalized activation by the function of squash (Zhao et al., 2018) to make the whole model nonlinear and get the predicted capsule, as shown in Equation 10 and 11.

$$s_j = \sum_i \theta_i^j \hat{fc}_{j|i} \tag{10}$$

$$Squash(s_j) = \frac{\|s_j\|^2}{1 + \|s_j\|^2} \frac{s_j}{\|s_j\|} \tag{11}$$

where s_j is the original output capsule vector, which is the predicted capsule value before the square function is activated.

We introduce the activation network based on the dynamic routing algorithm in the traditional capsule network. The obtained prediction capsule was input into the activation network of linear transformation and ReLU activation function, and the final output was obtained by connecting with the residual fc_j feature capsule. The formula is shown in 12 and 13.

$$AN(x) = \max(\mathbf{x}W_{+}c_1, 0)W_2 + c_2 \tag{12}$$

$$predict = fc_j + AN(pc_i) \tag{13}$$

In order to match the attention weight of multi head attention, we need to softmax the dynamic routing weight.

3.4 Model Interpretability

We use the principle of back propagation (Wang et al., 2020) to reach local interpretability. That is, a sample of mammography report to explain why the repors was diagnosed as benign or malignant breast cancer. The model defines two interpretable parameters: attention weight α_i^j and dynamic routing weight θ_i^j. Attention weight indicates whether attention right perform aimportant function in the formation of feature capsule. The calculation method is shown in Equation 5 above. The dynamic routing weight determines the higher level classification capsule to which the current feature capsule will output it. The formula is described in the Equation 9. As a classifier, the weight matrix of the traditional full-connection layer is fixed after training (Zhang et al., 2019), which is not conducive to interpretation. However, in our model, all layers are fully connected. The capsule network part and the multi attention part are interdependent, which has an important impact on the research of local interpretation based on back propagation. As shown in the figure 4, prediction capsules are divided into three categories $j \in [1, 3]$: benign, suspected malignant, and malignant breast cancer. By setting the parameter $P1$ of the capsule classification layer and the parameter $P2$ of the feature layer extracted by the multi-head self-attention.

P_1 The largest routing weight was selected from θ_1^j,θ_2^j,$\cdots$,θ_M^j, and the largest θ_i^j indicated that the ith feature capsule made the largest contribution to the jth prediction capsule, indicating that $P1$ the largest routing weight had an important impact on the prediction of benign breast cancer. Meanwhile, based on the principle of backpropagation, for $P1$ corresponding feature capsules, $P2$ phrase embedding w_i which contributed the most to the largest feature capsule was found in the multi-head self-attention according to the weight of attention α_1^i, α_2^i,$\cdots$,α_N^i. Then the keywords are searched in the breast examination report After one-dimensional convolution layer. When words are colored, it is important to predict the results. The model can be visualized in the report to provide simple and clear help for doctors.

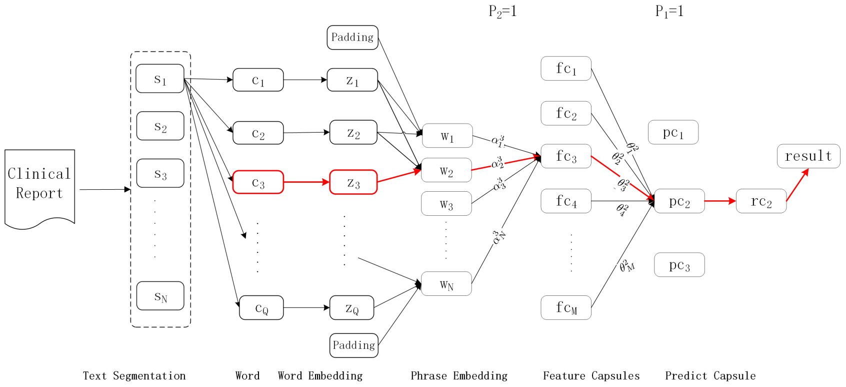


Figure 4. Model interpretability process display diagram.

4 Experiment

In this section, we mainly discuss the validity and interpretability of our model. Firstly, the selection of methods in different stages is considered, such as the performance differences of Sem-Bert, BERT and word2vec in acquiring word vectors. Secondly, the model in this paper is compared with some unexplainable text classification models such as TextCNN (Rakhlin, 2016) and LSTM (Kalchbrenner et al., 2015). Finally, we demonstrate the interpretability of our model through experiments.

4.1 Dataset

This dataset is the molybdenum mammography examination report of Shanghai Ruijin Hospital. We selected 1600 preoperative mammography data from 34 million original reports and included 2857 data after segmented pretreatment. The small sample dataset is classified into three categories, including benign, suspected malignant and malignant. The analysis of the data set is shown in Table 1.

Dataset	Classes	Benign	Suspected of Malignant	Malignant
Training Set	3	1344	390	268
Test Set	3	734	58	63
Mammography	3	2078	448	331

Table 1. Summary of mammography report dataset

4.2 Parameter Settings

The model was implemented by Tensorflow and trained by Adam Optimizer (Bock and Weiß , 2019). The multi-head attention part activation function uses a ReLU nonlinear function, the capsule network layer activation function Squash function. We set $P1 = P2 = 3$, the number of class capsules is 3, the number of head vectors is 18.

4.3 Different stage selection

Evaluation indexes: The main evaluation indexes of the experiment are Micro-Precision(Mi-P), Micro-Recall(Mi-R), Micro-F1-score(Mi-F1), Macro-Precision(Ma-P), Macro-Recall(Ma-R), Macro-F1-score(Ma-F1), Receiver Operating Characteristic(ROC).

Comparative experiment: The main purpose of this section is to verify the validity of the 'Sem-Bert + Muti-Cap' model, including the advantages of Sem-Bert in obtaining word vectors and the classification accuracy of Muti-cap. In this experiment, six other models are selected as the baseline. Comparing experiment 'word2vec+Muti-Cap' and 'BERT+Muti-Cap' in table 2 mainly verifies whether texts in specific fields such as medical care will affect the acquisition of word vectors, thus affecting the prediction performance. The comparative experiments textCNN, textRNN, LSTM, and Capsule in table 3 mainly verify the influence of different classification models on the prediction performance.

Experimental Results: Figure 5 shows the ROC curves of the model in this paper obtained according to different evaluation criteria under three classifications. For three classifications, the area under the curve and AUC values of micro and macro methods are different. In this paper, the dataset of benign, malignant and suspected malignant are large difference between three kinds of sample size, and there are obvious characteristics of malignant samples. So the ROC curve of class 2 malignant tag is more left, and the AUC area is larger. Meanwhile, the average (micro) AUC is larger than that of macro AUC. Since the samples in our dataset are unbalanced, so to give equal attention to the categories with small samples and those with large sample data. The ROC curves in subsequent comparative experiments were obtained by macro method.

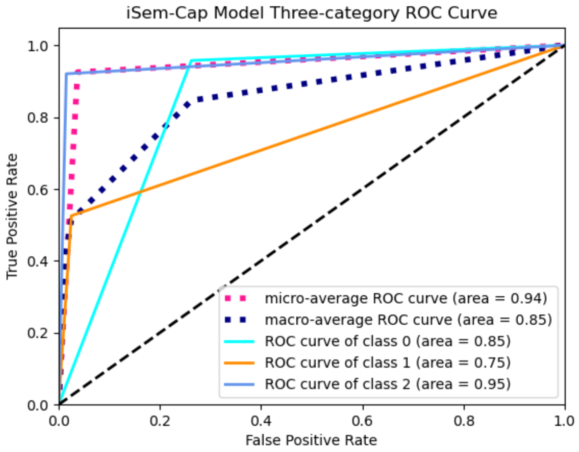


Figure 5. ROC curve of three classifications of BDCN model

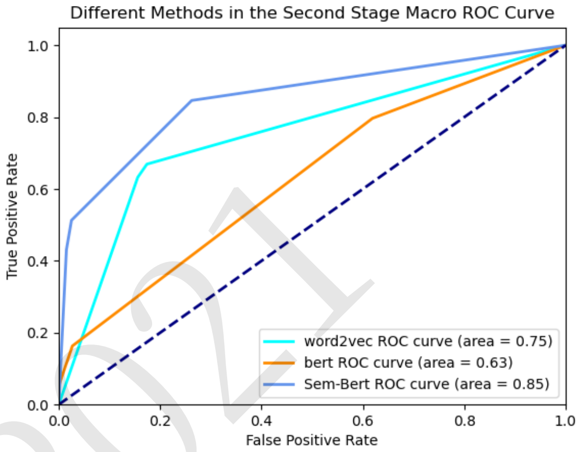


Figure 6. In the second stage, ROC curves of word vectors are obtained by different methods

Comparing (1) (2) with (3) in Table 2 and Figure 6, we can basically draw a conclusion: under the premise of using Muti-Cap method in the third stage, the evaluation index of the word vector method obtained by Sem-Bert in the second stage is higher, the ROC curve covers the other two comparative experiments, and the AUC value of the area under the ROC is significantly larger than the other two. Although there may be a loss of text information in the process of constructing semantic tree, the experiment effectively proves that the method of acquiring word vectors through semantic tree combined with BERT is more accurate than traditional word2vec and BERT method directly, which is very meaningful.

Model	**Evaluation index**					
	Mi-P(%)	**Mi-R(%)**	**Mi-F1(%)**	**Ma-P(%)**	**Ma-R(%)**	**Ma-F1(%)**
(1)word2vec+Muti-Cap	83.25	83.25	83.25	44.02	61.08	47.83
(2)BERT+Muti-Cap	87.37	87.37	87.37	78.26	42.37	45.56
(3)BDCN(our)	**91.58**	**91.58**	**91.58**	**75.95**	**79.73**	**77.14**

Table 2. Model selection at second stage

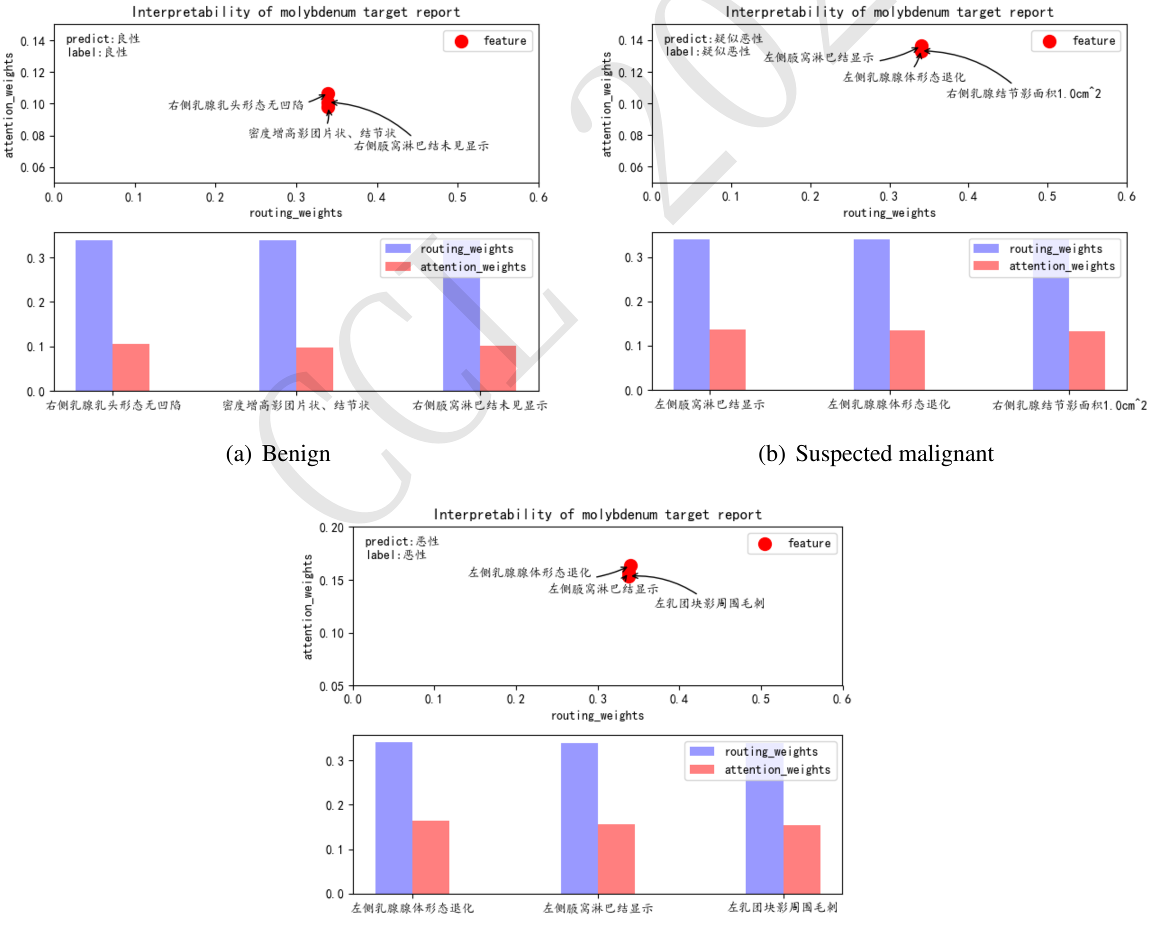
Table 3 compares the current mainstream deep learning models of text classification: CNN family textCNN, RNN family textRNN, LSTM family LSTM and capsule of Muti-Cap. The baselines are word2vec method to obtain the word vector, whereas in BDCN model, term vectors are obtained by Sem-Bert method. Comparing experiments (1), (2),(3) and (4) in table 5, the capsule network from the view on the text classification task to task, effect and less textCNN this classic mode, but better results than simple LSTM, which shows that the attempt is meaningful. Moreover, through the improvement of the traditional capsule network and word vector acquisition method, our model has favorable performance results in each index. Experiments show that the selection of the second stage of this model has greatly improved the accuracy of the model.

Model	Evaluation index					
	Mi-P(%)	Mi-R(%)	Mi-F1(%)	Ma-P(%)	Ma-R(%)	Ma-F1(%)
(1)textCNN	88.79	88.79	88.79	73.94	78.43	75.15
(2)textRNN	74.24	74.24	74.24	70.59	71.78	72.74
(3)LSTM	69.93	69.93	69.93	66.69	64.05	69.12
(4)Capsule	77.56	77.56	77.56	67.57	71.57	69.37
(5)BDCN(our)	**91.58**	**91.58**	**91.58**	**75.95**	**79.73**	**77.14**

Table 3. Comparison to multimodal baselines

4.4 Interpretability

Based on the back propagation principle, the BDCN can be interpreted to find the top three important weights in the dynamic routing weights according to the prediction results. Then the three important weights in the attention weights can be deduced in reverse, so as to find the three important features of judging the benign and malignant in this sample.



(a) Benign

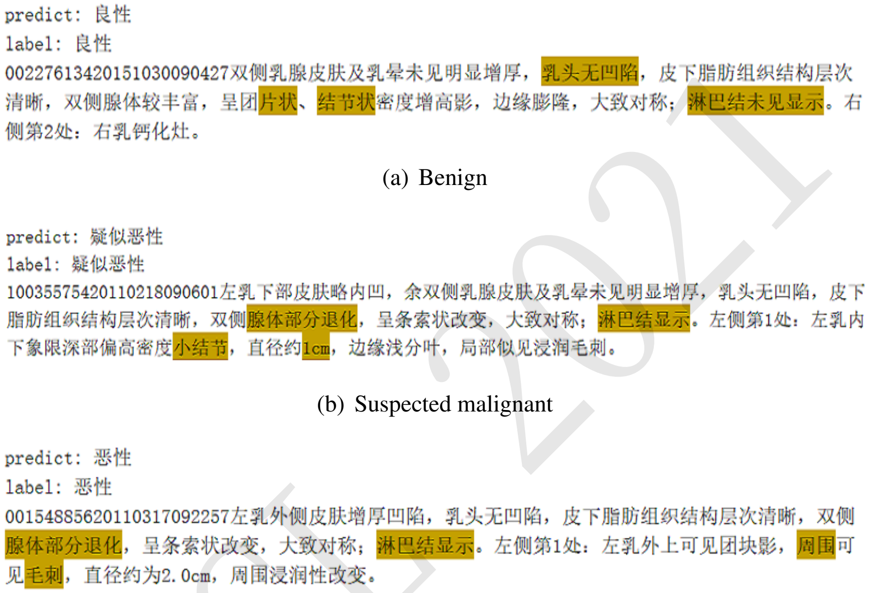
(b) Suspected malignant

(c) Malignant

Figure 7. Analysis of benign and malignant characteristics of breast molybdenum target

We plot a scatter plot and a bar chart, as shown in Figure 7. They respectively explained the important features of benign, suspected malignant and malignant cases as well as the attention weight and routing weight. For example, according to the ranking of importance, the three important features that appear for the first time of each sample were selected. In Figure (c), it can be seen that 'the shape of the burr', 'glandular degeneration' and 'lymph node display' are all important features of malignant tumors. In Figure (a) 'roughly symmetrical glands', 'density increase', 'no depression of the nipple' and 'no lymph nodes displayed' are all important indicators of benign breast tumors. Suspected malignancy is a grading result of BI-RADS4, the more similar the characteristics are to malignant, the greater the probability of malignant tumors, and it is convenient for the doctor to remind the patient for further examination.

More intuitively show the local interpretability of the model, the extracted important features were labeled and displayed on the segmented breast molybdenum target samples. As shown in Figure 8 (a) 'no inverted nipples', 'no lymph nodes' and 'increased density'. Benign and malignant breast tumors can be determined by directly observing the words marked yellow, and the prediction accuracy of the BDCN is also verified from the side. On the basis of concise and accurate text report, we add the annotation of important words to make up for the lack of interpretation of text classification better.



predict: 良性
label: 良性
00227613420151030090427双侧乳腺皮肤及乳晕未见明显增厚，乳头无凹陷，皮下脂肪组织结构层次清晰，双侧腺体较丰富，呈团片状、结节状密度增高影，边缘膨隆，大致对称；淋巴结未见显示。右侧第2处：右乳钙化灶。

(a) Benign

predict: 疑似恶性
label: 疑似恶性
10035575420110218090601左乳下部皮肤略内凹，余双侧乳腺皮肤及乳晕未见明显增厚，乳头无凹陷，皮下脂肪组织结构层次清晰，双侧腺体部分退化，呈条索状改变，大致对称；淋巴结显示。左侧第1处：左乳内下象限深部偏高密度小结节，直径约1cm，边缘浅分叶，局部似见浸润毛刺。

(b) Suspected malignant

predict: 恶性
label: 恶性
00154885620110317092257左乳外侧皮肤增厚凹陷，乳头无凹陷，皮下脂肪组织结构层次清晰，双侧腺体部分退化，呈条索状改变，大致对称；淋巴结显示。左侧第1处：左乳外上可见团块影，周围可见毛刺，直径约为2.0cm，周围浸润性改变。

(c) Malignant

Figure 8. Analysis of benign and malignant characteristics of breast molybdenum target

5 Conclusion

We proposed a semantically embedded self-interpreted breast diagnostic capsule network model. Semantic segmentation algorithm was used to segment the report, Sem-Bert method was used to obtain word vectors in medical field with hierarchical relationship, and capsule network with multiple attention was used to achieve prediction and classification of breast tumors. The validity of our model is better than other models in breast molybdenum target dataset. In addition, local self-interpretation method was used to provide intelligibility analysis, which was in line with doctors' clinical expectations. In the future, we will further study the global interpretability of the model and we hope to apply our technology to other diseases.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments. This work was financially supported by the National Key R&D Program of China under Grant 2019YFE0190500.

References

Aini H, and Haviluddin H. 2019. *Crude Palm Oil Prediction Based on Backpropagation Neural Network Approach*. Knowledge Engineering and Data Science, 2(1):1-9.

Bock S, and Weiß M. 2019. *A proof of local convergence for the Adam optimizer*. 2019 International Joint Conference on Neural Networks (IJCNN),Budapest,HU.

Chinese Anti-Cancer Association, Committee of Breast Cancer Society. 2019. *Chinese Anti-Cancer Association Guidelines and Specifications for Diagnosis and Treatment of Breast Cancer 2019 Edition*. Chinese Journal of Cancer,Shanghai,CHN.

Cerda-Mardini P, Araujo V, and Soto A. 2020. *Translating Natural Language Instructions for Behavioral Robot Navigation with a Multi-Head Attention Mechanism*. arXiv preprint arXiv:2006.00697.

Chen D, Huang M, and Li W. 2019. *Knowledge-powered deep breast tumor classification with multiple medical reports*. IEEE/ACM transactions on computational biology and bioinformatics,CA,USA.

Devlin J, Chang M W, Lee K, and Toutanova K. 2018. *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805.

Jiang D, and He J. 2020. *Tree Framework With BERT Word Embedding for the Recognition of Chinese Implicit Discourse Relations*. IEEE Access,8:162004-162011.

Kalchbrenner N, Danihelka I, and Graves A. 2015. *Grid long short-term memory*. arXiv preprint arXiv:1507.01526.

Liu W, Zhou P, Zhao Z, Wang Z, Ju Q, Deng H, and Wang P. 2019. *K-bert: Enabling language representation with knowledge graph*. In Proceedings of the AAAI Conference on Artificial Intelligence,34(3):2901-2908.

Long J, Shelhamer E, and Darrell T. 2015. *Fully convolutional networks for semantic segmentation*. Proceedings of the IEEE conference on computer vision and pattern recognition,3431-3440,Boston,MA.

Lundberg S, and Iwata S I. 2017. *A Unified Approach to Interpreting Model Predictions*, Advances in Neural Information Processing Systems,30: 4765-4774.

Mengwan W, Yongzhao D, Xiuming W, Qichen S, Jianqing Z, Lixin Z, Guorong L, and Jiafu Z. 2020. *A Benign and Malignant Breast Tumor Classification Method via Efficiently Combining Texture and Morphological Features on Ultrasound Images*,Volume 2020. Computational and Mathematical Methods in Medicine ,England,UK.

Mikolov T, Chen K, Corrado G, and Dean J. 2013. *Efficient estimation of word representations in vector space*, arXiv preprint arXiv:1301.3781.

Papineni K, Roukos S, Ward T, and Zhu W J. 2002. *Bleu: a method for automatic evaluation of machine translation*. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, Philadelphia,PA.

Rakhlin A. 2016. *Convolutional Neural Networks for Sentence Classification*. GitHub.

Ren Z, Hu X, and Ji S. 2019. *Evaluating Generalization Ability of Convolutional Neural Networks and Capsule Networks for Image Classification via Top-2 Classification*, arXiv preprint arXiv:1901.10112.

Ribeiro M T, Singh S, and Guestrin C. 2016. *Why should i trust you?" Explaining the predictions of any classifier*, Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining,1135-1144,New York, NY.

Sabour S, Frosst N, and Hinton G E. 2017. *Dynamic routing between capsules*, arXiv preprint arXiv:1710.09829.

Shrikumar A, Greenside P, and Kundaje A. 2017. *Learning Important Features Through Propagating Activation Differences*, International Conference on Machine Learning, PMLR,3145-3153.

Tang G, Müller M, Rios A, and Sennrich R. 2018. *Why self-attention? a targeted evaluation of neural machine translation architectures*. arXiv preprint arXiv:1808.08946.

Wang Z, Hu X, and Ji S. 2020. *iCapsNets: Towards Interpretable Capsule Networks for Text Classification*, arXiv preprint arXiv:2006.00075.

Wang Z, Li M, Wang H, Jiang H, Yao Y, Zhang H, and Xin J. 2020. *Breast cancer detection using extreme learning machine based on feature fusion with cnn deep features*. IEEE Access,1-1.

Wang L, Yang Y, and Nishikawa R M. 2009. *Microcalcification classification assisted by content-based image retrieval for breast cancer diagnosis*. Pattern recognition, 42(6):1126-1132.

Yoshikawa Y, and Iwata T. 2020. *Neural Generators of Sparse Local Linear Models for Achieving both Accuracy and Interpretability*, arXiv preprint arXiv:2003.06441.

Zhang W, Cai L, Chen M, and Wang N. 2019. *Progress in Interpretability Research of Convolutional Neural Networks*. In International Conference on Mobile Computing, Applications, and Services, Springer,Berlin,DE.

Zhao W, Ye J, Yang M, Lei Z, Zhang S, and Zhao Z. 2018. *Investigating capsule networks with dynamic routing for text classification*. arXiv preprint arXiv:1804.00538.

Zhao W, Peng H, Eger S, Cambria E, and Yang M. 2019. *Towards scalable and reliable capsule networks for challenging NLP applications*. arXiv preprint arXiv:1906.02829.