Shengyan at VQA-Med 2020: An Encoder-Decoder Model for Medical Domain Visual Question Answering Task

Shengyan Liu, Haiyan Ding, and Xiaobing Zhou*

School of Information Science and Engineering,
Yunnan University, Kunming 650091, P.R.China
Corresponding author: zhouxb@ynu.edu.cn

Abstract. Intelligent learning and understanding of image and text information are important research directions for the successful application of deep learning in computer vision (CV) and natural language processing (NLP). This paper takes medical images and questions as the research objects, by extracting the feature information contained in the medical images and questions and combining with the attention mechanism makes the computer system to more accurately obtain the information expressed by the images. Then, the model predicts the answers to the questions about the images. This paper proposes a novel model for the ImageCLEF VQA-Med 2020 task [1]. In this model, we use the improved pre-trained VGG16 to extract image features, and GRU module to extract text features of the questions. Then the structure of Seq2seq, including encoding and decoding parts, is applied to obtain the predicted answers. Our team gets the seventh rank in the ImageCLEF VQA-Med 2020 challenge, and our model achieves accuracy score and BLEU score of 0.376 and 0.412 respectively, in the competition.

Keywords: VQA-Med · VGG16 · Seq2seq · GRU · Attention Mechanism

1 Introduction

With the rapid development of CV and NLP, visual question answering (VQA) has become one of the increasingly popular research areas in deep learning [2]. VQA technology is a comprehensive technology that combines CV, natural language understanding, knowledge representation and reasoning. Compared with specific artificial intelligence technologies such as image processing, text processing, and NLP, VQA is a frontier for general artificial intelligence research explore [3]. Because it includes two parts of content about artificial intelligence, i.e., image processing and NLP. In the field of NLP, language-based question

answering has been extensively studied and great achievements have been made. However, question answering systems involving vision is rarely known. VQA is an interdisciplinary research direction. Its main purpose is to automatically answer natural language questions based on relevant visual content (pictures or videos). It is one of the key research directions in the field of artificial intelligence in the future. Most of the VQA technology is applied to some random scenes containing objects or people, and then generates questions related to the image based on the image content, finally, the VQA system gives the answers to the questions. In general, applying the attention mechanism or target detection (similar to Fast-RCNN [4]) method in the above-mentioned scenarios is more effective, but because the medical datasets lack labels and artificially divided candidate bounding boxes, there are no special pre-trained models on large medical datasets, so VQA tasks in the medical field are still difficult challenges. This paper describes the implementation of VQA in the medical field. The model we propose in this paper is a multi-classification one. The answers in the training set are extracted as candidate answer sets. The answers are divided into a simple yes/no, one word, and sentences composed of multiple words. The model uses CNN and RNN to extract the image and text features respectively, and uses these two parts of features as input to the next step, where the image is not the original one but a pre-processed one when it is input to the network. The purpose is to reduce noise in the image. The structure of this paper is organized as below.

The next chapter briefly describes the relevant work and summarizes the methods used in this model. Chapter 3 describes our proposed method and dataset. Chapter 4 introduces our model in detail. Chapter 5 describes the experimental results and model evaluation results, and Chapter 6 is the summary of this paper.

2 Related Work

After reading a lot of literature, we found that the implementation of VQA technology is generally based on deep learning, and deep neural networks are also the most effective method to achieve VQA tasks. VQA tasks are roughly divided into two aspects. First of all, CNN is generally used to extract image features, such as VGGNet, Resnet, Inception, Googlenet, and so on. Pre-trained deep learning networks by ImageNet [5] have obtained good results on many traditional VQA datasets, such as COCO-QA [6], Visual7W [7]. While we do not have a deep learning model pre-trained on large medical datasets, so we can only use the ImageNet pre-trained model and improve it. Secondly, RNN is used to extract the feature of the text processed through embedding layer. In this paper, the structure of seq2seq and the attention mechanism are applied. The attention mechanism first appeared in the Deepmind team using it to help classify images on the RNN model [8], and achieved good results. Subsequently, Bahdanau et al. [9] proposed the use of attention mechanism in machine translation tasks to complete machine translation and alignment work, which also achieved a huge

breakthrough. The sequence to sequence (Seq2seq) [10] method was proposed by the Google team in 2014. The basic idea of seq2seq is to use two RNNs, one RNN as the encoder, and the other as the decoder. We proposed an Xception-GRU model in the ImageCLEF VQA-Med 2019 task [11] last year, in which the Xception network was applied to the image feature extraction part, and the GRU model was applied to the text feature extraction part, these two parts of features were respectively passed through the attention module and the feature fusion module, and finally predicted answers after softmax layer. The model achieved accuracy and BLEU scores of 0.21 and 0.393 at ImageCLEF VQA-Med 2019 task and got the fifteen rank last year. Based on the new data set, we have made a little progress and gets the seventh rank in this year's ImageCLEF VQA-Med 2020 task. We will introduce this new data set, ImageCLEF VQA-Med 2020 dataset, in the following chapter 3 dataset description.

3 Dataset Description

The dataset in this paper is from the ImageCLEF VQA-Med 2020 task, which is divided into three parts, training set, validation set and test set as shown in Table 1. Compared to last year's data set, the pattern for this year's data set is one image for multiple questions instead of one image for one question last year, and this year's data set is not divided into four categories last year's data set, but the type of this year's questions is closer to last year's abnormality class questions. It is also the hardest class of questions to deal with, because the answers to the corresponding questions are not very regular.

Table 1. Statistics of VQA-Med data.

	Training	Validation	Test
Images	3000	500	500
Questions	3000	500	500
Answers	3000	500	—

In continuation of the two previous editions, this year's task on VQA-Med consists in answering natural language questions from the visual content of associated radiology images, it focuses particularly on questions about abnormalities [12].

There are two examples of medical images and associated questions and answers from the training set of ImageCLEF VQA-Med 2020, as shown in Figure 1:

Q: "what is most alarming about this ultrasound?"
A: "omphalocele"

Q: "what is most alarming about this ct scan?"
A: "hodgkin' s lymphoma (nodular lymphocytic predominant)"

Fig. 1. Two examples of medical images and associated questions and answers from the training set of ImageCLEF VQA-Med 2020.

4 Methods

4.1 Model prediction

This paper proposes an Encoder-Decoder model. The answers from the training set are extracted to form a candidate answer set. There are a total of 333 candidate answers. All we have to do is to let the model assign a predicted probability value to each answer word in this candidate answer set. The output module consists of GRU network that takes the thought vector which includes question and image features as initial state. $< SOS >$ token is taken as input in the first time step, then the GRU network tries to predict the answer using softmax layer. This method can be expressed as a mathematical formula:

$$y = \text{argmax}\, P(a|q, i, m), \tag{1}$$

where y is the candidate answer word option with the highest probability predicted by the model, q is the answer to the question, i is the image corresponding to the question, m provides all parameters of the model.

4.2 Sequence to sequence

The model we propose in this paper uses the sequence to sequence method. The general structure of this method is composed of an encoding module and a decoding module, as shown in the Figure 2:

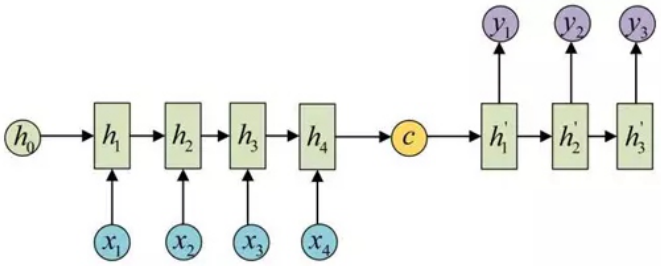

Fig. 2. The basic structure of encoding and decoding

The encoder is responsible for compressing the input sequence into a vector of a specified length. This vector can be regarded as the semantics of the sequence. This process is called encoding. As shown in the Figure 2, the simplest way to obtain the semantic vector is to directly use the hidden state of the last input as semantic vector C. It can perform a transformation on the last hidden state to obtain a semantic vector, and also perform a transformation on all the hidden states of the input sequence to obtain a semantic vector. The calculation formula is:

$$C = q(h_1, h_2, h_3, h_{tx}) = h_{tx}, \tag{2}$$

where h_i represents the output of each hidden layer, C is the state of the last input h_{tx}.

The decoder is responsible for generating the specific sequence based on the semantic vector. This process is called decoding. As shown in the Figure 2, the simplest way is to input the semantic variables obtained by the encoder as the initial state into the decoder's RNN to obtain the output sequence. It can be seen that the output of the previous moment will be used as the input of the current moment, and the semantic vector C only participates in the operation as the initial state, and the subsequent operations are independent of the semantic vector C. The calculation formula is:

$$y_i = g(y_{i-1}, h_i^{'}, C), \tag{3}$$

where y_{i-1} is the output of the previous step, $h_i^{'}$ is the output of the hidden layer, and g represents the nonlinear activation function.

The following symbols are represented as inputs at the decoding stage:

$< PAD >$: Complete characters.

$< EOS >$: End-of-sentence identifier on the decoder side.

$< UNK >$: Low-frequency words or some words have not encountered so on.

$< SOS >$: The start identifier of the sentence on the decoder side.

4.3 Implementation Details

Encoder In terms of image feature extraction in the encoding module, we use an improved VGG16 model [13] to extract image features, as shown in Figure 3:

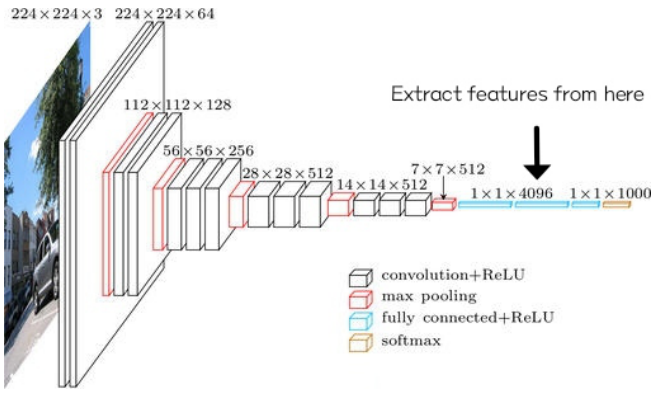

Fig. 3. The basic structure of VGG16 network

Extracting image features refers to inputting an image preprocessed in form of pixels into a feature vector with high-level semantic information. Convolutional neural networks as feature extractors are all standard models proposed in ImageNet image recognition tasks, and CNN models can be used to indirectly used a large amount of training data on ImageNet to perform better feature extraction on images. This paper uses the pre-training VGG-16 model as the visual feature extractor of the images. Since the last two layers have entered the classification step, we need the complete output image features, so the last two layers are removed, and the 4096-dimensional features are extracted from the fully connected layer, and then the output feature vector passes through an attention module [14]. Because the mapping relationship between the global features of the image and the sentences is not enough, it will bring a lot of noise signals. We need to extract the local features of the image, which requires us to use the attention mechanism to find the relationship between local image features. The basic unit of sentences can better complete the task from images to sentences so that images can be better combined with text features at the semantic level.

In terms of text feature extraction, we input the question text into RNN after Glove Embedding [15], and then summarize the output of each hidden layer to generate a semantic vector. GRU [16] is a variant of LSTM [17], which cancels the cell state in LSTM and only uses Hidden state, and use the update gate to replace the input gates and forget gate in the LSTM, cancel the output gate in the LSTM, and add the reset gate. The advantage of this structure is that under the premise of achieving similar effect of LSTM, the calculation on training is smaller, and the training speed is faster. Figure 4 is the structure of GRU model.

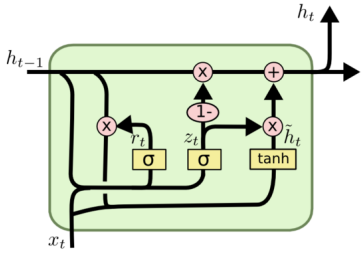

Fig. 4. The structure of GRU model

The forward propagation formula of GRU is as follows:

$$\begin{aligned} z_t &= \sigma(W_z \cdot [h_{t-1}, x_t]) \\ r_t &= \sigma(W_r \cdot [h_{t-1}, x_t]) \\ \widetilde{h_t} &= tanh(W \cdot [r_t * h_{t-1}, x_t]) \\ h_t &= (1 - z_t) * h_{t-1} + z_t * \widetilde{h_t} \end{aligned} \tag{4}$$

where z_t is the update gate, which is the logic gate when updating activation, r_t is the reset gate, whether to give up the previous activation h_t when deciding on candidate activation, $\widetilde{h_t}$ is candidate activation, receive $[x_t,h_{t-1}]$, h_t is activate gate, which is the hidden layer of GRU, receive $[h_{t-1},\widetilde{h_t}]$.

The following Figure 5 is the model structure of encoding part we used.

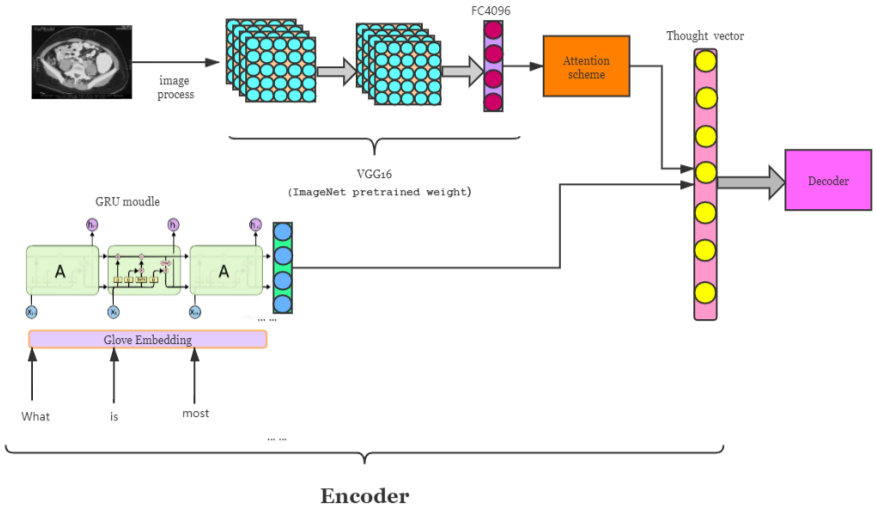

Fig. 5. Encoding part of the model

Decoder The semantic variables obtained by the encoder are input into the GRU of the decoder as the initial state to obtain the output sequence. The output of the previous moment will be used as the input of the current moment, and finally the predicted answer will be output.

The following is the model structure of decoding part we used.

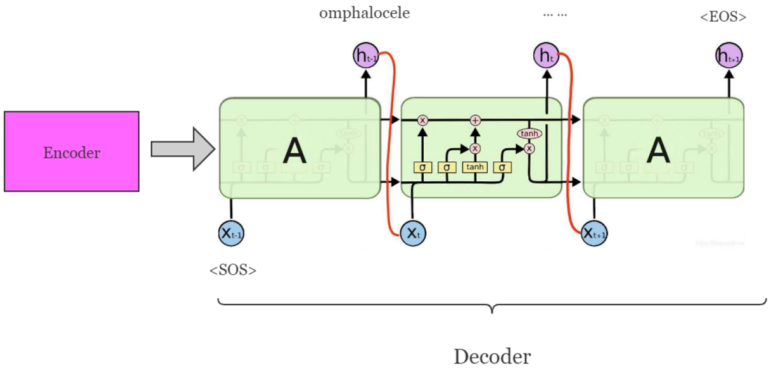

Fig. 6. Decoding part of the model

As shown in Figure 6, in this stacked GRU network, the red curve represents the hidden state information of the previous moment as the input of the next moment, the model input $< SOS >$ represents the start of decoding, the model prediction output $< EOS >$ represents the end of prediction. This is the decoder part of our model.

5 Evaluation and Result

ImageCLEF VQA-Med 2020 competition implements two evaluation methods, Accuracy (Strict) and BLEU [18]. It uses an adapted version of the accuracy metric from the general domain VQA task that considers exact matching of a participant provided answer and the ground truth answer and uses the BLEU metric to capture the similarity between a system-generated answer and the ground truth answer.

The implementation of the BLEU method is to calculate the N-grams model of the candidate sentence and the reference sentence [19]. Each answer is pre-processed in the following way: The caption is converted to lower-case; All punctuation is removed an the caption is tokenized into its individual words; Stopwords are removed using NLTK's "english" stopword list; Stemming is applied using NLTK's Snowball stemmer. The answer is always considered as a single sentence, even if it actually contains several sentences. [1] And then count the number of matches to calculate. This method has nothing to do with the

word order. Based on the model and method mentioned above, we submitted five results in the competition. The results of the competition have been shown in Table 2. Our team ID is "Shengyan".

Table 2. Official results of ImageCLEF VQA-Med 2020.

Participants	Accuracy	BLEU
z_liao	0.496	0.542
TheInceptionTea	0.480	0.511
bumjun_jung	0.466	0.502
going	0.426	0.462
NLM	0.400	0.441
harendrakv	0.378	0.439
Shengyan	**0.376**	**0.412**
kdevqa	0.314	0.350
sheerin	0.282	0.330
umassmednlp	0.220	0.340

As shown in the Table 3, the Xception+GRU model was proposed in last year's competition by our team. The traditional CNN and RNN models were used in image processing and text processing respectively, which did not perform very well in this year's dataset. This year, we mainly introduced the encoding and decoding structure of seq2seq and made ablation experiments based on last year's model. We can see that the model with seq2seq construct achieves better accuracy. The VGG16+GRU+seq2seq model proposed in this paper is improved based on the traditional CNN model to reduce the number of parameters and improve the accuracy. The hyperparameters of this model are set as follows: we set the learning rate to 0.0001 in ADAM optimizer, with dropout = 0.5, epoch = 80 and batchsize = 64. The following is a comparison of the results of all the experiments we performed. It can be seen that the VGG16-seq2seq model is the best in this paper.

Table 3. Results of our experiments on test set

model	Accuracy	BLEU
VGG16+GRU	0.28	0.35
Xception+GRU	0.21	0.39
Xception+GRU+seq2seq	0.30	0.40
GoogleNet+GRU+seq2seq	0.26	0.36
VGG16+LSTM+seq2seq	0.34	0.41
VGG16+GRU+seq2seq	**0.376**	**0.412**

6 Conclusion

This paper describes the model we use in the ImageCLEF VQA-Med 2020 competition. We use the seq2seq framework to input feature and predict answers. The image feature extraction part uses the improved VGG16 model. The text feature extraction uses the GRU model, and finally achieves the accuracy score of 0.376, and BLEU score of 0.412. We will improve the model and combine the attention mechanism in the Seq2seq structure to continuously improve the accuracy. Besides, our future work includes: (1) Image and natural language are signals of two modalities. How to fully integrate these two modalities belongs to the task of multi-modality fusion, which requires us to design a type that can fully learn the relationship between different modalities. (2) If the image visual features and the text features of the questions are directly fused, there will be a semantic level mismatch, so we will design a model to handle this question and improve the accuracy of VQA system.

Acknowledgments

This work was supported by the Natural Science Foundations of China under Grants 61463050, the NSF of Yunnan Province under Grant 2015FB113.

References

1. Bogdan Ionescu, Henning Müller, Renaud Péteri, Asma Ben Abacha, Vivek Datla, Sadid A. Hasan, Dina Demner-Fushman, Serge Kozlovski, Vitali Liauchuk, Yashin Dicente Cid, Vassili Kovalev, Obioma Pelka, Christoph M. Friedrich, Alba García Seco de Herrera, Van-Tu Ninh, Tu-Khiem Le, Liting Zhou, Luca Piras, Michael Riegler, Pål Halvorsen, Minh-Triet Tran, Mathias Lux, Cathal Gurrin, Duc-Tien Dang-Nguyen, Jon Chamberlain, Adrian Clark, Antonio Campello, Dimitri Fichou, Raul Berari, Paul Brie, Mihai Dogariu, Liviu Daniel Ştefan, and Mihai Gabriel Constantin. Overview of the ImageCLEF 2020: Multimedia retrieval in lifelogging, medical, nature, and internet applications. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, volume 12260 of *Proceedings of the 11th International Conference of the CLEF Association (CLEF 2020)*, Thessaloniki, Greece, September 22-25 2020. LNCS Lecture Notes in Computer Science, Springer.
2. Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. 2016.
3. Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Vqa: Visual question answering. 2017.
4. Ross Girshick. Fast r-cnn. *Computer Science*, 2015.
5. Jia Deng, Wei Dong, Richard Socher, Li Jia Li, and Fei Fei Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, 2009.

6. Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. 2015.
7. Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. 2015.
8. Lan Lin, Huan Luo, Renjie Huang, and Mao Ye. Recurrent models of visual co-attention for person re-identification. *IEEE Access*, pages 1–1, 2019.
9. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *Computer Science*, 2014.
10. Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 2014.
11. Asma Ben Abacha, Sadid A. Hasan, Vivek V. Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. VQA-Med: Overview of the medical visual question answering task at imageclef 2019. In *CLEF 2019 Working Notes*, CEUR Workshop Proceedings, Lugano, Switzerland, September 09-12 2019. CEUR-WS.org <http://ceur-ws.org>;.
12. Asma Ben Abacha, Vivek V. Datla, Sadid A. Hasan, Dina Demner-Fushman, and Henning Müller. Overview of the vqa-med task at imageclef 2020: Visual question answering and generation in the medical domain. In *CLEF 2020 Working Notes*, CEUR Workshop Proceedings, Thessaloniki, Greece, September 22-25 2020. CEUR-WS.org.
13. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *Computer Science*, 2014.
14. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *Computer Science*, pages 2048–2057, 2015.
15. Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.
16. Kyunghyun Cho, Bart Van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *Computer Science*, 2014.
17. S Hochreiter and J Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
18. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.
19. Jessica Perrie, Aminul Islam, Evangelos Milios, and Vlado Keselj. Using google n-grams to expand word-emotion association lexicon. In *Proceedings of the 14th international conference on Computational Linguistics and Intelligent Text Processing - Volume 2*, 2013.