An Xception-GRU Model for Visual Question Answering in the Medical Domain

Shengyan Liu, Xiaozhi Ou, Jiao Che, Xiaobing Zhou(✉) and Haiyan Ding

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

Abstract. This paper introduces an Xception-GRU model for ImageCLEF 2019 Medical Domain Visual Question Answering (VQA-Med) Task. First, we enhance the images and remove extraneous words from the questions and convert the questions to vectors. Then, we employ pretrained Xception model to extract image features and use GRU model to encode the questions. To generate the output, we combine these two models with the attention mechanism. Our Xception-GRU model achieves the accuracy score of 0.21 and BLEU score of 0.393.

Keywords: VQA-Med · Xception · GRU · Attention Mechanism

1 Introduction

With the extensive application of deep learning in Computer Vision (CV) and Natural Language Processing (NLP), the powerful feature learning ability of deep learning greatly promotes the research in CV and NLP. Various deep networks represented by Convolutional Neural Network (CNN) emerge endlessly in CV, which can learn image features end-to-end without relying on the features of manual design. Through feature extraction layer by layer, CNN combines images from simple edges, corners, and other low-level features into higher-level features layer by layer. CNN's powerful feature extraction ability makes it more efficient to extract and compress image information. Recurrent Neural Network (RNN) model also shows its power in the field of NLP, especially in speech recognition, machine translation, language model and text generation. Visual Question Answer (VQA) consists of CV and NLP content, which inputs an image and an arbitrary form of natural language question about the image, and finally outputs a natural language answer. Medical VQA can help doctors improve their confidence in diagnosis and help patients better understand their conditions through the automatic system. In this paper, we propose an Xception-GRU model for ImageCLEF Medical Visual Question Answering (Med-VQA)

2019 Task [2] [7]. The model takes an image and a question as an input and outputs the answer to this question based on features combined both image and question features with an attention mechanism.

The rest of this paper is structured as follows. The next section will provide a brief overview of the work involved. The dataset provided in the rang of the VQA-Med challenge is described in Section 3. The deep learning networks that we proposed for VQA in the medical domain is presented in Section 4. The submitted runs is described in Section 5. Finally, Section 6 is the summary of the paper.

2 Related Work

Because VQA involves the two domains of CV and NLP, a natural VQA solution is to combine CNN and RNN, which have been very successful in CV and NLP, respectively, to construct a combination model. The VQA model, which is composed of deep CNN and LSTM network structure, is a relatively good model in visual question answering. Among them, some of the superior VQA models are introduced below.

Deeper LSTM Q + norm I model [1]. This model is proposed by Aishwarya Agrawal et al. In which, "I" refers to the extracted image features, and "norm I" refers to L2 normalization of image semantic information vector (1024 dimension) extracted by CNN. CNN extracts image semantic information and LSTM extracts text semantic information contained in the problem, and then the two information are fused so that the model can learn the meaning of the problem. Finally, the answer output is generated in a multi-layer MLP with Softmax as the output layer.

VIS+LSTM model [11]. This model is proposed by Mengye Ren et al. The basic structure of the model is to extract image information with CNN at first, and then connect LSTM to generate prediction results.

Neural-Image-QA model [9]. This model is proposed by Mateusz Malinowski et al. Based on CNN and LSTM, a model with length-variable prediction result is designed. In this model, visual question answering task is regarded as an auxiliary sequence to sequence task combined with image information.

mQA model [5]. This model is proposed by Gao H et al. In their paper, the understanding of visual question and answer task is that this model needs to give an answer to the question of the free form of an image, and the answer can be a sentence, a phrase or a word. The mQA model consists of four sub-modules, the first module encodes natural statements into a dense word vector feature by a LSTM network, i.e., extracts the information contained in the problem, called the problem LSTM network; The second module extracts image features from a deep CNN; The third module is another different LSTM network, which is used to code the characteristic information of the current word and some previous words in the answer, called the answer LSTM network; The last module fuses the information generated by the previous three models to predict the next word to be generated in the answer.

Most of the work discussed in this section cannot be directly applied to the VQA-Med for two reasons. The first one is obvious, this task mainly focuses on the medical domain, which gives this problem its unique set of challenges. As for the other one, it is related to how the sentences of the answers are constructed in VQA-Med, which is different from existing VQA datasets, such as DAtaset for QUestion Answering on Realworld images (DAQUAR) [8], Visual7W [14], Visual Madlibs [13], COCO-QA [11], Freestyle Multilingual Image Question Answering dataset (FM-IQA) [5], Visual Question Answering (VQA) [1], etc.

3 Dataset Description

This dataset of ImageCLEF 2019 VQA-Med [2] differs from previous data sets in that it divides the problems into four categories based on modality, plane, organ system, and abnormality. The purpose is to generate a more focused set of problems for the evaluation of the results. The dataset contains 12,792 QA pairs, 3,200 medical images for training sets, and 2,000 QA pairs, 500 medical images for validation sets, and 500 medical images with 500 questions for test sets. Fig.1 shows an example of a medical image and the associated question and answer from the training set of VQA-Med 2019 dataset. Table 1 lists the statistics of VQA-Med 2019 dataset.

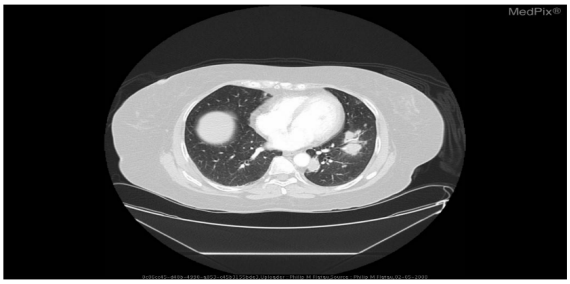


Fig. 1. An example of a medical image and the associated question and answer from the training set of ImageCLEF 2019 VQA-Med.

Examples of these four categories are shown below:

1. Modality, e.g. What kind of image is this? Was IV contrast given to the patient?
2. Plane, e.g. What plane is the image acquired in? In what plane is this image oriented?

3. Organ System, e.g. What organ system is primarily present in this image? What organ system is shown in this CT scan?
4. Abnormality, e.g. What is abnormal in the CT scan? What abnormality is seen in the image?

Table 1. Statistics of VQA-Med 2019 dataset.

	Training	Validation	Test
Images	3200	500	500
Questions	12792	2000	500
Answers	12792	2000	—

4 The Xception-GRU Model

Although at present the research in the VQA field has made some achievements, there's still a challenging problem that the overall accuracy of the answer by using existing models to realize the visual question and answer the problem is not high. The existing VQA models are relatively simple in structure, the content and form of answers are relatively simple, and more priori knowledge is needed for slightly complex problems, so simple reasoning cannot make correct answers. The reason is that, in addition to the image information of CNN, the knowledge source of LSTM in the learning process is only focused on the training question and answer pairs, with simple knowledge structure and lack of information. After comparing the characteristics of each pre-training CNN model, we propose the following model for our participation in VQA-Med 2019.

4.1 The Main Model

The model we proposed is as Fig.2:

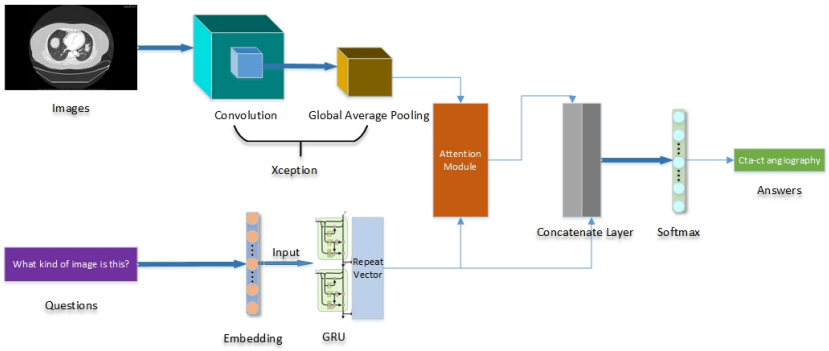


Fig. 2. The architecture of Xception-GRU module

We use Xception [4] to extract image features and GRU to extract question features. Since the number of image features is much larger than question features, the question features should be repeated to make the number of the two features equal, and then an attention mechanism [12] is added before fusion.

4.2 Image Representation

In recent years, many CNN models have been proposed, such as AlexNet, VGGNet, Inception, Xception, ResNet, etc. In this paper, we use Xception to extract image features. Xception was proposed by Francois Chollet, author of Keras, in 2017. It is another improvement of Inception-v3 proposed by Google after Inception. The advantage of Xception is that it can improve the efficiency of network, as well as in the case of a number of equal participation. On large data sets, the effect is better than Inception-v3. This also provides another idea of "lightweight": increasing network efficiency and performance as much as possible in the case of given hardware resources, which can also be understood as making full use of hardware resources. The architecture of Xception is shown in Fig.3.

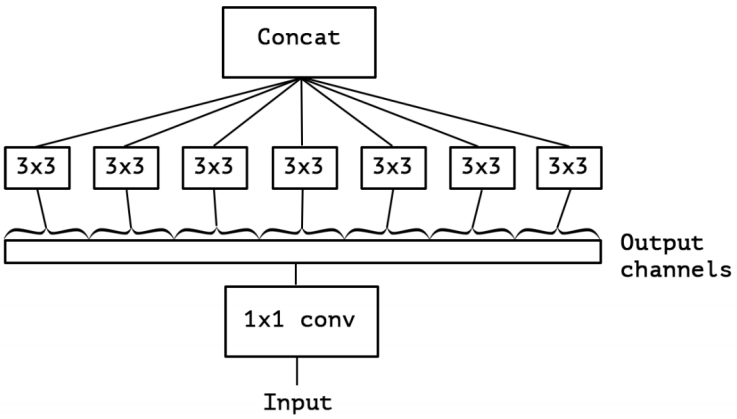


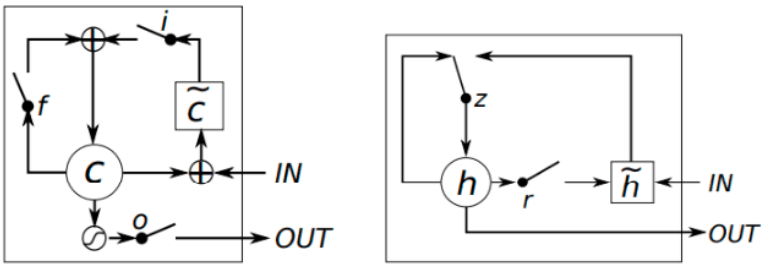
Fig. 3. The architecture of Xception module

First, the image goes through a convolution kernel of 1*1, the function of 1*1 convolution is to reduce dimension, and because each convolution kernel convolves only with the corresponding channel, the network uses separate convolution kernels. Finally, the features of each channel are joined together. The benefit about this method is that we get features that are independent of each other, without too much redundancy.

4.3 Question Representation

Gated Recurrent Unit(GRU) [3] was proposed by Cho van Merrienboer, Bahdanau and Bengio in 2014. By introducing the concept of gate, the calculation

method of hidden state in the cyclic neural network is modified, which includes reset gate, update gate, candidate hidden state and hidden state. GRU is one gate less than LSTM. We can think of the GRU as an optimization or variation of the LSTM. Resetting the gate helps capture short-term dependencies in the time series. Update gates help capture long-term dependencies in time series, but the experimental results are quite similar:



(a) Long Short-Term Memory (b) Gated Recurrent Unit

Fig. 4. illustration of (a) LSTM and (b) gated recurrent units. (a) i, f and o are the input, forgetand output gates, respectively. c and $\widetilde{c}$ denote the memory cell and the new memory cell content. (b) r and z are the reset and update gates, and h and $\widetilde{h}$ are the activation and the candidate activation.

When we train a GRU network, the input of the output layer is:

$$y_t^i = W_0 h.$$

The output is:

$$y_t^o = \sigma(y_t^i).$$

The loss function at a certain moment is:

$$E_t = \frac{1}{2}(y_d - y_t^o)^2.$$

In this part, we use the GRU model to extract the features of questions after preprocessing them.

5 Evaluation and Results

Before running the evaluation metrics on the answers, the following preprocessing are performed:

1. The capitals in each answer are converted to lowercases,
2. All punctuations are deleted and each answer is tokenized by single words.

The evaluation can be conducted based on the following metrics:

Accuracy (Strict) Accuracy is our most common evaluation indicator, and it is easy to understand. For a given test dataset, the number of correct samples in the test task is divided by the number of all samples. Generally speaking, the higher the accuracy, the better the classifier.

BLEU How to measure the similarity between the generated statement and the reference statement is an important issue. In 2002, Kishore Papineni et al proposed a classic measure standard BiLingual Evaluation Understudy (BLEU) [10]. BLEU is an auxiliary method to evaluate the quality of Bilingual translation. This method is simple, short, fast and easy to understand. Because the effect is reasonable, it has been widely migrated to various assessment tasks of natural language processing. It is used to determine how similar machine-translated sentences are to human-translated sentences. BLEU calculates the ratio of similarity between two sentences by counting the frequency of words appearing together, using the n-gram matching rule. BLEU evaluations are fast and close to human ratings. So the performance of a VQA model can be judged with the BLEU score. The higher the score, the better performance of a VQA model.

Three experiments are conducted to evaluate our model. The parameters are set as follows, the size of dictionary is 1000, the length of sequences is 9, the hidden size of GRU is 128, and the batch size of training is 256. We set the epoch to 54.

The experiments are described as follows.

1. In the first experiment, we run our proposed model (Xception-GRU) without date enhancement.
2. In the second experiment, we use Bi-LSTM instead of GRU to extract text features. Obviously, bidirectional LSTM is less effective than GRU.
3. In the last experiment, we run our proposed model (Xception-GRU) with date enhancement, the rest of the architecture stays the same.

The following table shows the results obtained on the test set:

Table 2. Results of our proposed model on Test set.

Model	Accuracy	BLEU
Xception + GRU **without** enhancement	0.21	0.393
Xception + Bi-LSTM **without** enhancement	0.2	0.31
Xception + GRU **with** enhancement	0.178	0.27

As shown in Table 2, our proposed Xception-GRU model without data enhancement achieves good results in term of BLEU metric (0.393) and accuracy (0.21). When we try Bi-LSTM model [6] to extract question features and Xception to extract image features without data enhancement, the effect is reduced a

little. Then we remain Xception-GRU architecture, and introduce data enhancement, the effect is reduced a lot. The reason is that due to the high performance of Xception model and the depth of feature extraction, it is easy to overfit. Therefore, Xception without image enhancement produces better results, while with image enhancement produces worse results.

In this regard, we still need to make some improvements on the mechanism to prevent overfitting. However, since there are no medical imaging professionals who can provide suggestions for the improvement of our process, the results may differ from the actual situation.

6 Conclusion

In this paper, we present our contribution to the visual question answering task in the field of medicine in view of the very meaningful but challenging VQAMed Task of ImageCLEF 2019. Our Xception-GRU model achieves the accuracy score of 0.21 and BLEU score of 0.393.

Our future work will focus on making the answers more readable and accurate. We consider that there is an essential semantic gap between the regional visual features and the source of the problem text representation. Due to the great success of the attention-based model in VQA task [12], we want to work on features, namely how to extract visual information more effectively and apply the attention mechanism better. We will improve our visual model by using attention feature enhancement techniques to further make regional semantic representations more relevant to the problem. Our future work also includes training on multiple data sets, improving model performance, etc.

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. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
2. Ben Abacha, A., Hasan, S.A., Datla, V.V., Liu, J., Demner-Fushman, D., Müller, H.: VQA-Med: Overview of the medical visual question answering task at imageclef 2019. In: CLEF 2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org/Vol-2380/.>;, Lugano, Switzerland (2019)
3. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
4. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1251–1258 (2017)

5. Gao, H., Mao, J., Jie, Z., Huang, Z., Lei, W., Wei, X.: Are you talking to a machine? dataset and methods for multilingual image question answering. Computer Science pp. 2296–2304 (2015)
6. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural Networks **18**(5-6), 602–610 (2005)
7. Ionescu, B., Müller, H., Péteri, R., Cid, Y.D., Liauchuk, V., Kovalev, V., Klimuk, D., Tarasau, A., Ben Abacha, A., Hasan, S.A., Datla, V., Liu, J., Demner-Fushman, D., Dang-Nguyen, D.T., Piras, L., Riegler, M., Tran, M.T., Lux, M., Gurrin, C., Pelka, O., Friedrich, C.M., de Herrera, A.G.S., Garcia, N., Kavallieratou, E., del Blanco, C.R., Rodríguez, C.C., Vasillopoulos, N., Karampidis, K., Chamberlain, J., Clark, A., Campello, A.: ImageCLEF 2019: Multimedia retrieval in medicine, lifelogging, security and nature. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 10th International Conference of the CLEF Association (CLEF 2019), LNCS Lecture Notes in Computer Science, Springer, Lugano, Switzerland (2019)
8. Malinowski, M., Fritz, M.: A multi-world approach to question answering about real-world scenes based on uncertain input (2014)
9. Malinowski, M., Rohrbach, M., Fritz, M.: Ask your neurons: A neural-based approach to answering questions about images. In: IEEE International Conference on Computer Vision (2015)
10. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. pp. 311–318. Association for Computational Linguistics (2002)
11. Ren, M., Kiros, R., Zemel, R.: Exploring models and data for image question answering. In: International Conference on Neural Information Processing Systems (2015)
12. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. arXiv preprint arXiv:1502.03044 (2015)
13. Yu, L., Park, E., Berg, A.C., Berg, T.L.: Visual madlibs: Fill in the blank image generation and question answering. In: IEEE International Conference on Computer Vision (2016)
14. Zhu, Y., Groth, O., Bernstein, M., Li, F.F.: Visual7w: Grounded question answering in images (2015)