Employing Inception-Resnet-v2 and Bi-LSTM for Medical Domain Visual Question Answering

Yangyang Zhou, Xin Kang, and Fuji Ren

Tokushima University, Tokushima 770-8506, JP
c501737062@tokushima-u.ac.jp
{kang-xin, ren}@is.tokushima-u.ac.jp

Abstract. In this paper, we describe our method for generating the answers for questions based on medical images, in the ImageCLEF VQA-Med 2018 task [7][5]. Firstly, we use some image enhancement methods like clipping and questions preprocessing methods like lemmatization. Secondly, we use Inception-Resnet-v2 model (CNN) to extract image features, and use Bi-LSTM model (RNN) to encode the questions. Finally, we concatenate the coded questions with the image features to generate the answers. Our result was ranked secondly based on the BLEU, WBSS and CBSS metrics for evaluating semantic similarity, which suggests that our method is effective for generating answers from medical images and related questions.

Keywords: VQA-Med · Inception-Resnet-v2 · Bi-LSTM · Attention mechanism.

CEUR-WS.org/Vol-2125/paper_107.pdf

1 Introduction

Visual question answering (VQA) is the task of generating textual answers for questions based on the contents of images. The VQA system takes images and questions as input, and combines the information of the input to generate readable answers as output. To generate the answers of specific questions, the VQA system needs to understand the content of the images and to get related background knowledge, which involves natural language processing and computer vision techniques. On the other hand, with the increasing of pouring attention into the medical domain, the combination of VQA and medical domain has become an extremely interesting challenge. It can not only provide a reference of diagnosis to the doctor, but also allow the patient to obtain health information directly, thereby improving the efficiency of diagnosis and treatment. Existing systems like MYCIN [13] have been able to simulate the diagnostic process and generate treatment plans based on relevant medical knowledge and a series of rules.

This paper aims to generate readable answers in the ImageCLEF VQA-Med 2018 task. The dataset involves a variety of medical images, related questions and answers. We divide the data into two parts as input. We use some image enhancement methods, and generate the image features by pre-trained CNN model. As for the part of the questions, we use kinds of text preprocessing methods like lemmatization, and after that, encode the questions by RNN model. Then, we add attention mechanism to the model. At last, we formulate simple rules on the output and generate reliable answers.

The rest of this paper is organized as follows. Section 2 briefly reviews the related work of VQA-Med task. Section 3 describes the analysis of data sets and methods used for generation in details during the experiment. We report our experiment result and evaluation in section 4, and conclude this paper in section 5.

2 Related work

A very close study to the VQA-Med task is the VQA challenge[1]. The VQA challenge has been held every year since 2016. The data set is based on open domain and includes more than 260 thousand images and 5.4 questions on average per image.

Kafle K et al. [8] and other researchers summarized quite a few methods for VQA. The majority of them used recurrent neural networks such as LSTM to encode questions, and used deep convolutional neural networks such as VGG-16 to focus on image recognition in advance. On the basis of these, there were variant models such as attention mechanisms [17], neural modules [1], dynamic memory [10], and even the addition of external knowledge bases [16], to improve the accuracy of the answers.

Deep convolutional neural networks [9] (CNN) can be used to extract the features of an image and identify the objects in it. The Inception-Resnet-v2 model [15] is one kind of advanced convolutional neural network that combines the inception module with ResNet. The remaining connections allow shortcuts in the model to make the network more efficient.

Elman J L [3] first used a recurrent neural network (RNN) to handle sequences problems. Nevertheless, context information is easily ignored when RN-N processes long sequences. The proposal of LSTM [6] alleviated the problem of long-distance dependence. Furthermore, the researchers also found that if the input sequence is reversed, the corresponding path from the decoder to the encoder will be shortened, contributing to network memory. The Bi-LSTM model [4] combines the two points above, and makes the result better.

On the other hand, there have been many Computer-aided diagnosis systems in medical imaging [2]. However, the majority of them are dealing with single-disease problems, and mainly concentrated on easily-determined regions such as the lungs and skins. The progress of the complex parts is slow. Compared with detection technology, the global lesions and structural lesions are still intensely difficult for the machines to learn.

The VQA-Med task differs from the VQA challenge in that it requires the understanding of different kinds of medical images with different body parts.

[1] http://visualqa.org/index.html

3 Methodology

3.1 Dataset analysis

	Training	Validation	Test
Images	2278	324	264
Questions	5413	500	500
Answers	5413	500	

Table 1. Statistics of VQA-Med data

The dataset of VQA-Med task consists of more than two thousand images, containing several kinds of medical images, such as computed tomography, magnetic resonance imaging, positron emission tomography, etc. However, compared to the open field VQA dataset, the number of training examples in the VQA-Med task is very small. For the deep learning models of VQA, which usually contain millions of parameters, the learning process would converge quickly with high bias, i.e. overfitting. Table 1 shows the statistics of the data. From the training set, there are an average of 2.4 questions per image, and a maximum of 7 questions per image. This ratio is even smaller in the validation set and test set. Additionally, there is only one reference answer for each question, which has a great limitation for answers generation.

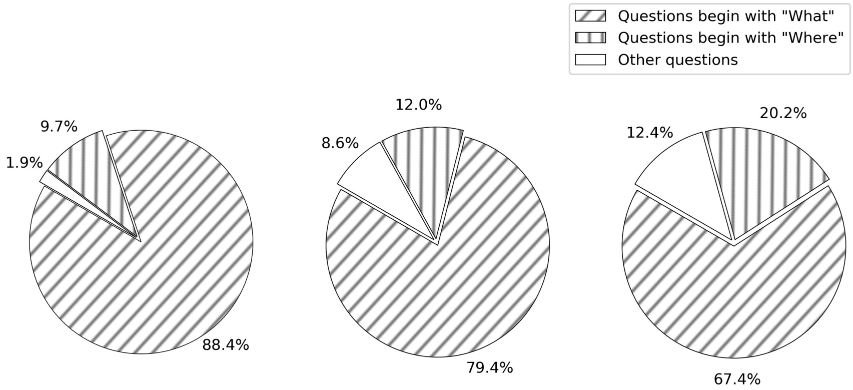

Fig. 1. Distribution of VQA-Med question types

Fig.1 shows statistics about different types of questions. The number of questions started with word “what” is large in three datasets, while the questions asking positions and other questions, including Yes-no questions, occupy relatively small proportions. Moreover, the proportions of questions in the three datasets are also quite different. Therefore, it is difficult for computers to learn the characteristics from the questions in small proportion in the course of training, and the performance in validation and test may not be as good as we expected.

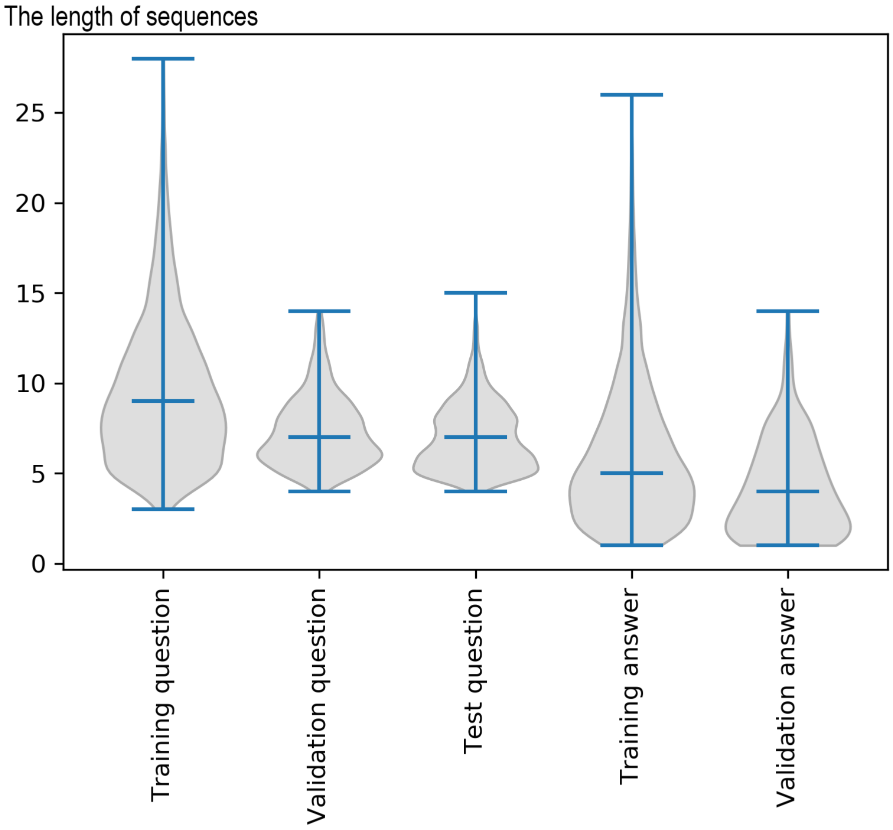

Fig. 2. Distribution of VQA-Med text sequences

We count the sequences of questions and answers. As shown in Fig.2, the sequence length of the questions is obviously longer than the answers. In fact, many of the answers are just phrases and cannot form complete sentences. Through horizontal comparison, we can find that the length of sentences in the training set is longer than that of the validation set. To prevent too slow training due to the long sequences, and to prevent loss of information due to the short sequences, we fix the length of the training sentences to 9 words. This length of sentences allows us to reserve the contents of most of the questions and answers.

Specifically, the "empty" words will be filled up at the end of the short sentences, and only the beginning 9 words will be reserved for the long sentences.

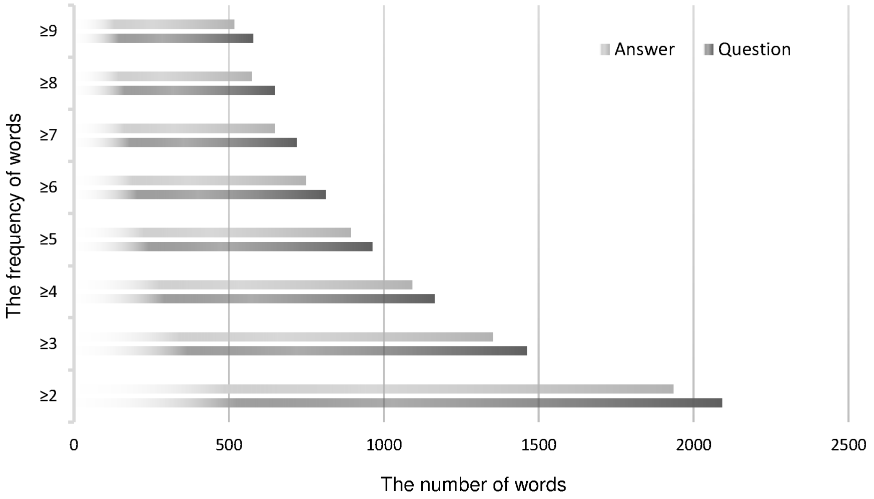

Fig. 3. Distribution of VQA-Med text Word Frequency

After merging all the questions and answers separately, we calculate the word frequency in Fig.3. In order to ensure the effectiveness of training, we plan to remove low-frequency words. Considering that it is appropriate to control the dictionary size of questions or answers within one thousand, we eventually set the words whose frequency is less than 5 as low-frequency words.

3.2 Preprocessing

For images, we use Inception-Resnet-v2 models to generate their features. In order to reduce the overfitting case, we adopt some image enhancement methods. Considering there are position judgments in the task, we reconstruct the picture with exceedingly small random rotations, offsets, scaling, clipping, and increase to 20 images per image (Fig.4).

For questions, we adopt some methods like stemming and lemmatization to alter verbs, nouns, and other words into original forms, to prevent overfitting. Furthermore, there is a situation that both full name and abbreviation coexist, like "inferior vena cava" and "IVC". We have changed all these medical terms into abbreviation. There are also a lot of pure numbers and combinations of numbers and letters. Therefore, the combinations of letters and numbers used to represent positions are mapped to an "pos" token, and the pure numbers are mapped to an "num" token, so as to reduce information complexity.

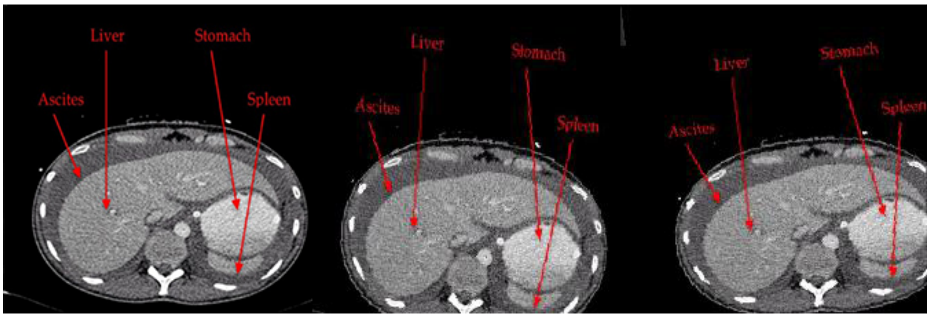

Fig. 4. An example of original (left) and enhanced (middle and right) images

In addition, we try to remove useless information such as stop words. According to the word frequency distribution in data analysis, we remove the low-frequency words to ensure training efficiency. In the meanwhile, we establish the dictionaries separately and make sure that the sizes of the dictionaries are both within one thousand.

There are some high-frequency verbs like "show" that emerge in almost every question. Several less useful adjectives like "large" also appear in questions from time to time. To cooperate with image enhancement methods, these verbs and adjectives are removed in the questions each time, so that each question is enhanced to 20 questions, and the answer remains unchanged at the same time.

The preprocessing of the answers is simpler than that of the questions. We use lemmatization and removing stop words. Besides, we create dictionaries separately and make sure that the sizes of them are within one thousand, just like questions part. However, the difference is that the low-frequency words in answers would be replaced by "abnormality" instead of simply removing them. Words with numbers have not been replaced. And the output sequences are the same as input sequences.

3.3 VQA-Med model

The basic model we build is to combine Inception-Resnet-v2 with Bi-LSTM. Firstly, as shown in Fig.5, the medical images are transformed into the features through the Inception-Resnet-v2 network. The pre-training weights of the Inception-Resnet-v2 are based on the Apache License[2]. Secondly, the questions are fed to the embedding layer and the Bi-LSTM layer. The last time step output of the Bi-LSTM layer is reserved as the question encoding features. Thirdly, after concatenating the features of images and questions, we use another Bi-LSTM layer. And this time the output returns decoded sequences. Finally the fully connected layer outputs the predicted sequence with "softmax" activation.

[2] https://github.com/tensorflow/models/blob/master/LICENSE

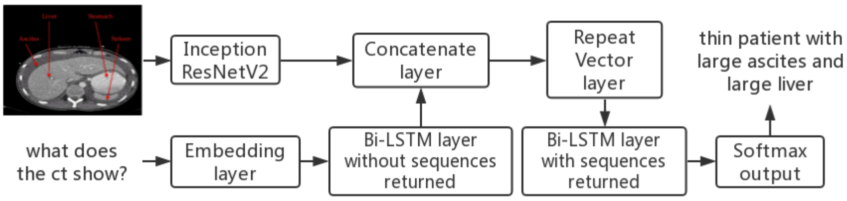
Inception ResNetV2
Concatenate layer
Repeat Vector layer
thin patient with large ascites and large liver
what does the ct show?
Embedding layer
Bi-LSTM layer without sequences returned
Bi-LSTM layer with sequences returned
Softmax output

Fig. 5. Initial model without attention mechanism

The loss function of the model we selected is categorical cross entropy, using the following formula:

$$H(T,q) = -\sum_{i=1}^{n} \frac{1}{N} \log_2 q(x_i) \tag{1}$$

where N is the size of validation set, and q(x) is the probability of event x estimated from the training set.

Considering that the overfitting is severe with small amount of training data, we adopt a dropout value of 50%, a L2 regularization in the Bi-LSTM layers and a batch normalization after the Bi-LSTM layers. However, we find that there are some problems with the syntax and semantics of the generated answers, which is not satisfactory. In particular, the overfitting problem still exists.

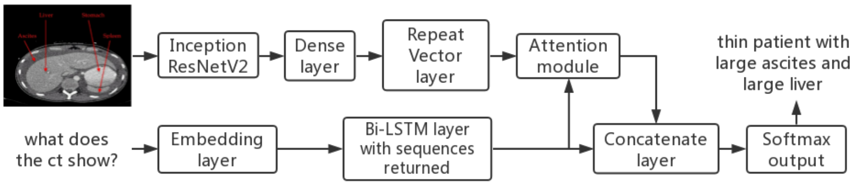

Fig. 6. Improved model with attention mechanism

To solve this problem, we add attention mechanism and modify the model. As shown in Fig.6, the image features are converted by Inception-Resnet-v2 network, as in the previous model. Then, we use a dense layer and a repeat vector layer to deal with the image features. The questions are trained in a Bi-LSTM layer after an embedding layer and return the sequences directly, with a 50% dropout rate and a batch normalization. We adopt the attention module to integrate the features of the images and the questions. After that, we concatenate the outcomes of attention module with the question features. Eventually, the full-connected layer outputs the predicted sequences with "softmax" activation.

We also added several simple rules to the output to make the generated answers more reasonable. It may be due to the fact that the word frequency of prepositions is relatively high, some of the generated answers have successive and repetitive prepositions outputs. Thus, we choose to delete these extra prepositions. In addition, for the answers of Yes-no questions, there is a case in which "yes" or "no" is output at the same time with other words unrelated. We choose to delete these extra words as well.

4 Experiment

4.1 Model selection

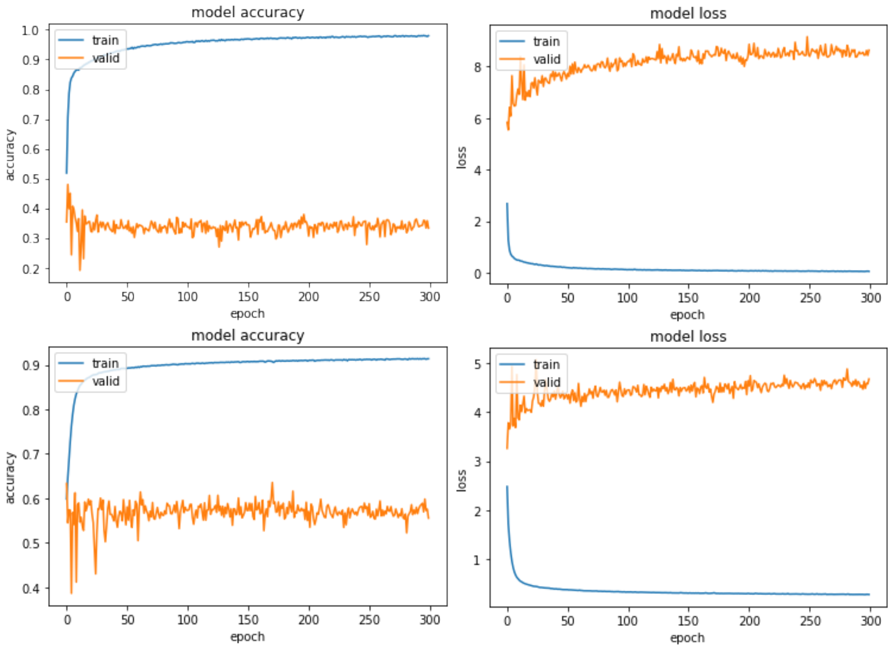

Fig. 7. Accuracy and loss of initial model(top); Accuracy and loss of improved model (bottom)

Based on the performance of VQA-Med on the validation set, the parameters are set as follows. The size of dictionary is 1000, and the length of sequences is 9. The hidden size of Bi-LSTM is 128. And the batch size of training is 256.

The metrics method is categorical accuracy. We use the ADAM optimizer with $\beta 1 = 0.9, \beta 2 = 0.999, \varepsilon = 10^{-8}$.

We set the epoch to 300, and the training process is shown in Fig.7. The accuracy of the validation set and the degree of overfitting are both better than that without attention mechanism. The final loss of no-attention-model is over 8 while that of attention-model is about 4.5, which means it is effective to add attention mechanism.

The final result we submitted is using training set and the validation set to participate in the training.

4.2 Evaluation

The following evaluation methods are employed for evaluating the VQA-Med results.

Bilingual evaluation understudy [12] is an auxiliary tool for assessing the quality of bilingual translations. It is used to determine the degree of similarity between sentences translated by machines and by humans. BLEU uses the matching rule of N-gram to calculate the proportion of similarity between two sentences. Actually, it is to calculate the frequency of two sentences co-occurrence words. This tool is fast, and the results are also close to human evaluation scores. Nevertheless, there are also deficiencies. For instance, it is easily interfered by frequent words, cannot consider synonym expression, and do not consider grammatical accuracy. In this task, the method is used to compare the similarity between the generated answers and the referenced answers.

Word-based Semantic Similarity method is used to measure the semantics similarity between the generated answers and the factual answers at the word level by tokenizing predictions and real answers as words. This algorithm is recently used to calculate the semantic similarity in the biomedical domain [14].

Concept-based Semantic Similarity is similar to WBSS as described above. The difference is that this metric is to extract the biomedical concepts in the predictions and the real answers respectively, then construct a dictionary. After vectorizing the words and calculating the cosine between them, the similarity could be expressed.

	BLEU	WBSS	CBSS
Improved model without output rules	0.103070853	0.147733901	0.3236155
Basic model with output rules	0.106454315	0.159756011	0.334431201
Improved model with output rules	0.134830654	0.173731936	0.329503441

Table 2. Scores of VQA-Med task submissions

The scores of our task submissions are shown in Table 2. As can be seen from the table, the use of output rules is crucial: the scores of all three evaluation methods drop if it is removed. Attention mechanism is also a significant component that improves BLEU and WBSS scores.

Most of the generated results are phrases, such as “right region” and “anterior part bladder”. 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.

5 Conclusion

In this paper, we described our participation in ImageCLEF VQA-Med 2018 task, which is a problem of answering questions for the medical domain. We use images and questions enhancement preprocessing. We adopt the VQA-Med model introduced above during the training. Our result has a BLEU score of 0.135, a WBSS score of 0.174 and a CBSS score of 0.330. As can be seen, due to the small number of datasets, it is difficult to generate highly accurate answers without using external data.

Our future work will focus on making the answers more accurate. In the preprocessing section, we can classify the medical images and train them separately. External data and relevant medical knowledge can be used in data enhancement. As for the model, we consider to use other new methods, such as Hierarchical Co-Attention model [11] to improve the accuracy of answers.

Acknowledgements

This research has been partially supported by the Ministry of Education, Science, Sports and Culture of Japan, Grantin-Aid for Scientific Research(A), 15H01712.

References

1. Andreas, J., Rohrbach, M., Darrell, T., Klein, D.: Neural module networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 39–48 (2016)
2. Doi, K.: Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Computerized medical imaging and graphics **31**(4-5), 198–211 (2007)
3. Elman, J.L.: Finding structure in time. Cognitive science **14**(2), 179–211 (1990)
4. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural Networks **18**(5-6), 602–610 (2005)
5. Hasan, S.A., Ling, Y., Farri, O., Liu, J., Lungren, M., Müller, H.: Overview of the ImageCLEF 2018 medical domain visual question answering task. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Avignon, France (September 10-14 2018)
6. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)

7. Ionescu, B., Müller, H., Villegas, M., de Herrera, A.G.S., Eickhoff, C., Andrearczyk, V., Cid, Y.D., Liauchuk, V., Kovalev, V., Hasan, S.A., Ling, Y., Farri, O., Liu, J., Lungren, M., Dang-Nguyen, D.T., Piras, L., Riegler, M., Zhou, L., Lux, M., Gurrin, C.: Overview of ImageCLEF 2018: Challenges, datasets and evaluation. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018), LNCS Lecture Notes in Computer Science, Springer, Avignon, France (September 10-14 2018)
8. Kafle, K., Kanan, C.: Visual question answering: Datasets, algorithms, and future challenges. Computer Vision and Image Understanding **163**, 3–20 (2017)
9. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
10. Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., Zhong, V., Paulus, R., Socher, R.: Ask me anything: Dynamic memory networks for natural language processing. In: International Conference on Machine Learning. pp. 1378–1387 (2016)
11. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. In: Advances In Neural Information Processing Systems. pp. 289–297 (2016)
12. 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)
13. Shortliffe, E.: Computer-based medical consultations: MYCIN, vol. 2. Elsevier (2012)
14. Soğancıoğlu, G., Öztürk, H., Özgür, A.: Biosses: a semantic sentence similarity estimation system for the biomedical domain. Bioinformatics **33**(14), i49–i58 (2017)
15. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI. vol. 4, p. 12 (2017)
16. Wang, P., Wu, Q., Shen, C., Hengel, A.v.d., Dick, A.: Explicit knowledge-based reasoning for visual question answering. arXiv preprint arXiv:1511.02570 (2015)
17. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning. pp. 2048–2057 (2015)