TUA1 at ImageCLEF 2019 VQA-Med: A classification and generation model based on transfer learning

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 article, we describe a method for answering questions based on medical images in the ImageCLEF VQA-Med 2019 task. The VQA-Med 2019 mission involves four categories: modality, plane, organ system, and abnormality. We try to turn the strong artificial intelligence problem into multiple weak artificial intelligence problems. First, we use a simple classifier to distinguish the four categories by training questions only. Transfer learning is useful in easing the overfitting problem of neural network on small datasets. Then, we use Inception-ResNet-V2 and Bidirectional Encoder Representation from Transformers pre-training model as feature extractors to deal with medical images and questions, respectively. Next, we use additional classifiers to get answers in the modality, plane and organ system categories. Last, we use sequence-to-sequence model as a generator to get the answer in the abnormality category. Our submission ranks fourth, based on accuracy metric and BLEU metric, which shows that our method can effectively get answers from medical images and related questions.

Keywords: VQA-Med · Inception-Resnet-v2 · BERT · seq2seq.

1 Introduction

VQA-Med2019 [**?**] [**?**] is a visual question answering (VQA) task in the medical domain. Health has consistently been our concern. Artificial intelligence (AI) has made significant breakthroughs in different tasks such as lesion recognition. Compared with image recognition, VQA task needs to further understand the content in images, which is more difficult. The VQA-Med2019 covers 36 image modalities including CT, MR, etc., 10 separate organ systems, 16 planes, and various abnormalities. Some examples are provided in Fig.1. Most medical datasets are created for a single condition, like lung cancer. And VQA-Med looks at a variety of conditions and hopes to help physicians and patients in diagnosis by establishing a question and answer system.

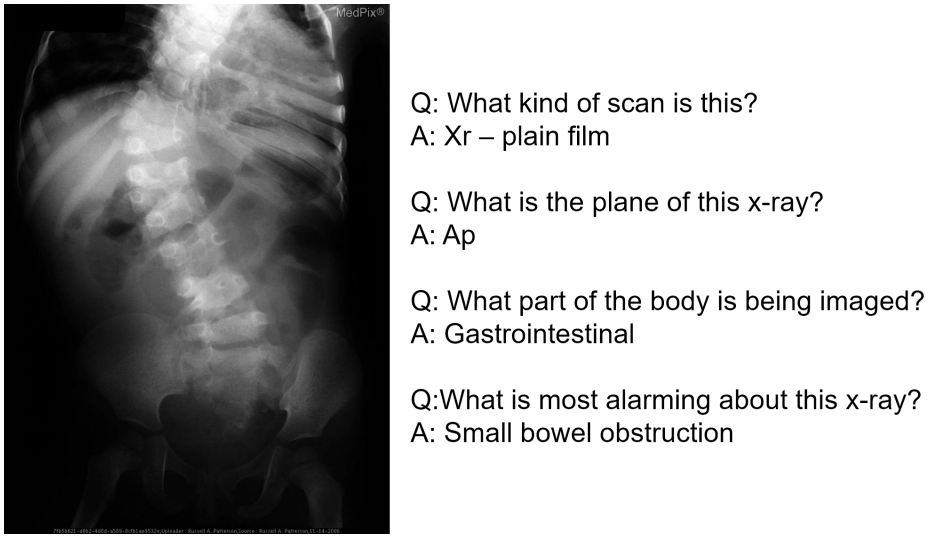

Fig. 1. VQA-Med2019 contains four categories of questions: modality, plane, organ system, and abnormality.

Most open domain VQA tasks choose the best answer from the alternative by classification. VQA-Med2019 adds the classification part to the VQA-Med2018 [**?**]. We use a simple classifier for questions only to break down this difficult task into multiple relatively easy tasks. In the modality, plane and organ system parts, we use the additional classifiers to obtain the answers, and in the abnormality part we use the generator to get the answers. Due to the small amount of VQA-Med2019 data, we choose to use transfer learning to prevent overfitting. Additional classifiers and generator use the Inception-ResNet-V2 [**?**] (IRV2) and Bidirectional Encoder Representation from Transformers [**?**] (BERT) pre-training model as feature extractors for processing images and questions, respectively. The generator is constructed based on a sequence-to-sequence [**?**] model.

The result we submitted gets 0.606 accuracy score, and 0.633 BLEU score, ranking fourth, which shows that our method can effectively get answers from medical images and related questions.

The rest of this paper is structured as follows. Section 2 briefly reviews the work related to the VQA-Med task. Section 3 provides details of the method we proposed. Section 4 reports our experimental results and evaluation results. Section 5 presents conclusions and future work.

2 Related work

In this task, we are using the VQA approach. VQA tasks involve image processing and natural language processing. The open domain VQA tasks have

various question objects and ways. Most VQA tasks are classified tasks. The existing method of processing VQA task is mainly the Hierarchical Co-Attention Model [**?**], which has a good performance in classification.

Some hospitals have begun using computer-aided diagnostic tools to advise physicians. However, Most of the existing auxiliary diagnoses are only for a certain field, such as lung cancer. As for the question and answering, due to the diversity of answers, the evaluation is quite difficult. So far the accuracy rate is far from the level of human physicians. In addition, datasets in the medical field, especially datasets involving multiple diseases, are rare, and medical image datasets of many research institutions are not disclosed.

The VQA-Med2019 data set involves a variety of conditions, but the existing technology cannot deal with different types of pictures such as MR, CT, and different organ parts such as brain, lungs, and abdomen at the same time. We try to build classifiers and generator to simplify this strong AI problem.

3 Methodology

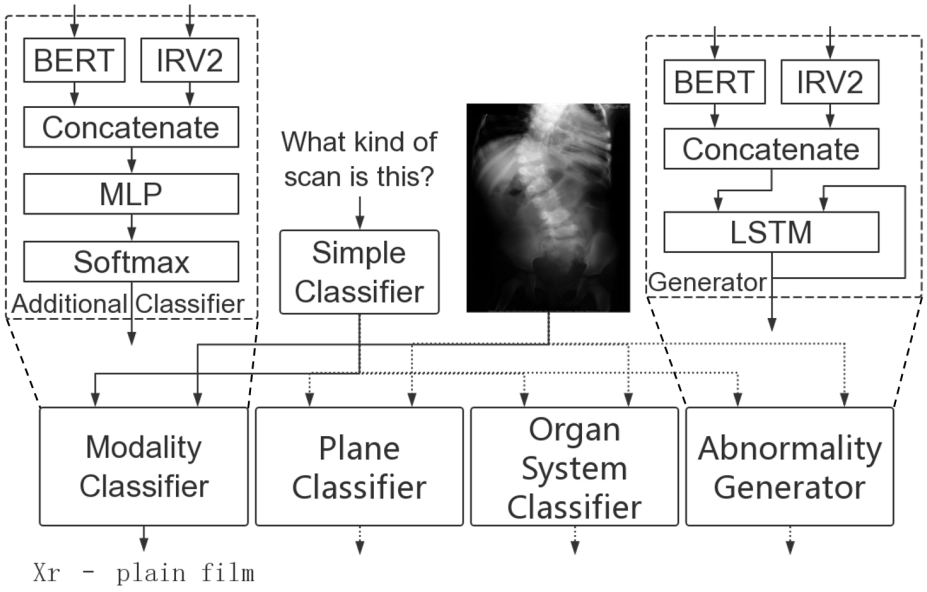

Fig. 2. An overview of our proposed model architecture which has four classifiers and a generator.

As mentioned in section 1, we propose a model for VQA-Med2019. This system is primarily composed of multiple classifiers and a generator. As is shown in Fig.2, simple classifier is used to classify questions, additional classifiers are

used for modality, plane and organ system questions, generator correspond to abnormality problems.

3.1 Preprocessing

For medical images, we use data enhancement methods to randomly shift, shear, and scale the image. We have not used methods such as rotation and inversion because there may be position-related questions. As for the texts, we convert all the questions and answers into lowercase letters and remove the punctuation.

3.2 Classifier

Since the test set does not provide the categories, we first train a simple classifier to classify questions. The simple classifier can divide the dataset into four distinct categories: modality, plane, organ system, and abnormality by observing only the questions. On this basis, we train three additional classifiers for modality, plane, and organ system problems, respectively. The data part of the next section describes the answers to these three types of questions are under a variety of options. Three additional classifiers extract features from images and questions: the features of the medical images are extracted by the IRV2 pre-trained on the Imagenet [**?**]; the features of the questions are extracted by the weighted public BERT pre-training model. After that, we use Multi-Layer Perception (MLP) to unify the feature dimensions of the images and questions and concatenate them together. Then, we use the classification layer to output the categories.

3.3 Generator

The generator is utilized to solve abnormality questions. There are not any options for this category, and all the answers need to be generated by the model from understanding the information of the images. The generator is built based on sequence-to-sequence (seq2seq) model. The encoder part, similar to the classifier, uses IRV2 to extract features of the medical images and concatenates the questions features extracted by the BERT as the initial state of the decoder. The decoder bases on a long-short term memory [**?**] network. We put "sos" token as the initial input, continuously loops to generate the probability distribution of the next word, and puts the most likely word as the input of the next moment until the output is "eos" token. In the prediction part, we use beam search [**?**] to output the answers.

4 Experiment

4.1 Dataset

The VQA-Med2019 training set contains 3200 medical images and clearly identifies four categories of questions and answers (Q&A) pairs of 12792 pairs; the

validation set contains 500 images and 2000 Q&A pairs; the test set contains 500 images and 500 questions, and no label for the type of questions. For modality, plane, organ system, the dataset gives the options for the answers (refer to the readme document in the dataset). As for the abnormality part, the answer involves a variety of specific disease types, and there are no options for us to choose. This dataset involves a wide range of medical images and Q&A pairs, close to the real medical environment.

4.2 Evaluation metrics

We choose the accuracy metric commonly used in VQA task. The accuracy metric considers the exact match between the answer and the fact.Besides, we choose BLEU [**?**] metric commonly used in machine translation. BLEU metric calculates the frequency of words that appear together between the answer and the fact, in other words, the similarity between the answer and the fact. Both metrics are automatically scored.

4.3 Result

We submit just one run. The result we submitted gets 0.606 accuracy score, and 0.633 BLEU score. This result ranks fourth in the leaderboard. Compare with the first ranking, we are 0.018 score behind in accuracy and 0.011 score behind in BLEU.

After comparing with the published ground truth, we found that in the categories of plane and organ system, both classification accuracies exceed 70%, but in modality, the accuracy is less than 70%. It may be due to the large number of modality options and the small difference between them (there are 17 sub-options of MR, such as "MR-STIR", "MR-FIESTA", etc.), which is difficult to distinguish.

As for the category of abnormality, the result we generated is quite different from the ground truth. Perhaps because of the small amount of data, the process of answer generation is much more related to the frequency of words in the training set (such as "multiforme", "glioblastoma", etc.), rather than the medical images.

5 Conclusion

In this article, we describe our participation in VQA-Med2019 Task: to answer questions based on the medical images. We use a simple classifier to categorize the questions, then use the additional classifiers or generator for different types of questions to get the corresponding answers. In this process, we use transfer learning including IRV2 and BERT to prevent over-fitting problem that might have on small dataset. We obtain 0.606 accuracy score, and 0.633 BLEU score, indicating that our proposed method has a certain effect. It may be explained

by the wide range of datasets and the small number of training for a particular condition. This method is far from being a human physician.

Our future work will focus on making the answers more accurate. Consider that there is a certain amount of noise in the dataset, such as Some CT images labelled as x-ray images. We may improve the accuracy of the answers by correcting or eliminating the noise.

Acknowledgements

This research has been partially supported by JSPS KAKENHI Grant Number 19K20345.

References

1. 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>, Lugano, Switzerland (September 9-12 2019)
2. 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)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
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., Péteri, R., Cid, Y.D., Liauchuk, V., Kovalev, V., Klimuk, D., Tarasau, A., Abacha, A.B., 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 Tenth International Conference of the CLEF Association (CLEF 2019), LNCS Lecture Notes in Computer Science, Springer, Lugano, Switzerland (September 09-12 2019)
8. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering (2016)
9. 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)

10. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
11. Wiseman, S., Rush, A.M.: Sequence-to-sequence learning as beam-search optimization. arXiv preprint arXiv:1606.02960 (2016)