kdevqa at VQA-Med 2020: focusing on GLU-based classification

Hideo Umada[1] and Masaki Aono[2]

Department of Computer Science and Engineering,
Toyohashi University of Technology, Aichi, Japan
1. umada@kde.cs.tut.ac.jp
2. aono@tut.jp

Abstract. Interpretation of medical images is a challenging research problem with increasing interest in medical applications of artificial intelligence. In particular, the ImageCLEF2020 visual question answering (VQA) task is expected to have applications such as a second opinion. The purpose of this research is to find an effective VQA-Med system method. We propose neural networks using the Gated Linear Unit for effective fusion of image and question features. Before training, we perform pre-processes and conduct pre-training. We apply so called "inpainting" to remove a logo or text embedded in images so that we attempt to extract image features with less noise. And we use the VQA-Med2019 dataset to train some of the weights of the proposed model. We consider the VQA task as a 332-dimensional classification task. The score of our proposed model turns out to be 0.314 in Accuracy and 0.350 in Bleu in VQA-Med2020 task.

Keywords: VQA-Med · Visual Question Answering · Classification · Inpainting.

1 Introduction

With increasing interest in artificial intelligence to support clinical decision-making and to improve patient engagement, the application to automated medical image interpretation is currently getting much popularity. In particular, it is expected that the second opinion provided by the automated system will enhance the judgment of clinicians.

Visual Question-Answering (VQA) is the task to generate a plausible answer presented with an image-question pairs such as left of Fig. 1. The task requires expertise in both natural language processing (NLP) and computer vision (CV) so that researchers have been attempting to solve the problem from various standpoints with Deep Neural Networks (DNN).

In this paper, we describe our approach to ImageCLEF2020 [1] visual question answering (VQA) task [2] in medical domain at VQA such as right of Fig. 1. The nature of medical images are quite different from general images such as Imagenet [3] in many aspects. The knowledge on medical vocabulary seems to the must to better understand both the questions and answers written in medical terminologies.

In the following, we first describe related work on VQA task and VQA-Med task in Section 2, followed by the description of the dataset provided for VQA-Med2020 dataset in Section 3. In Section 4, we describe details of the method we propose, and then of our experiments we have conducted in Section 5. We finally conclude this paper in Section 6.

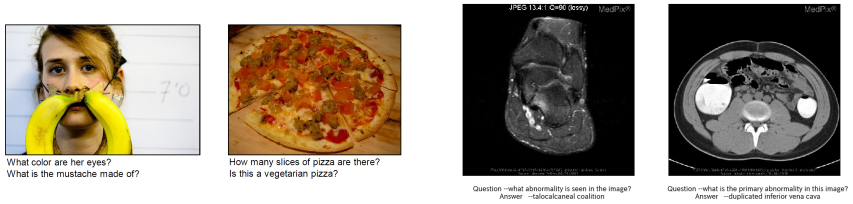

Fig. 1. Example of general (left) and medical (right) VQA data

2 Related Work

Convolution Neural Networks (CNNs) for image recognition, such as VGG and ResNet, has been used extensively. Similarly, multiple Transformers for sentence comprehension, such as BERT, has been getting popular recently. Accordingly, feature extraction from pretrained neural network models, transfer learning and fine tuning with the pretrained models have been actively investigated. Visual Question Answering, or VQA, stands between image recognition and sentence comprehension, and is regarded as a bridge application between them. Research on VQA is actively carried out through the VQA Challenge using VQA v2.0 [4]. For example, P.Anderson et al. proposed DNN using Bottom-up Attention [5] obtained by using pretrained Faster R-CNN [6] which is one of CNN used for object detection. In addition, as a VQA-Med task, there are competitions at ImageCLEF2018 and 2019. Yan et al. [7] proposed dividing the dataset into subcategories and attempted to solve the tasks by transforming the opriginal problem into a classification problem with categories in VQA-Med2019.

3 Dataset of VQA-Med2020

The VQA-Med2020 dataset consists of 5,000 pairs of medical image and question-answering. Specifically, the dataset consists of 4,000 training, 500 validation, and

500 test data. Most of the images in the VQA-Med2020 dataset are non-colored, and they potentially include non-essential logos and texts. The question pattern can be classified into 39 different types for training and validation data. In our analysis, the top 10 patterns cover more than 94% of the total data. On the other hand, there are 332 different answer patterns, and the top 10 patterns cover approximately 12% of the total data. Table 1 summarizes top 5 frequent questions and answers.

Table 1. Frequently Questions and Answers Ranking in VQA-Med2020

Rank	Question	freq	Answer	freq
1	what abnormality is seen in the image?	1,106	pulmonary embolism	88
2	what is the primary abnormality in ...	1,073	acute appendicitis	80
3	what is most alarming about this ct ...	482	angiomyolipoma	49
4	what is abnormal in the ct scan?	460	yes	49
5	what is abnormal in the mri?	252	adenocarcinoma of the lung	46

4 Proposed method

This section presents our methods in VQA-Med2020. The overview of our system is illustrated in Fig 2 with the yellow layers having trainable weights. We deal with VQA as a classification task of 332-dimension. All the images make some pre-processes shown in subsection 4.1 and are later characterized by VGG. We use VGG16 with batch normalization model [8] pretrained at Imagenet [3] to extract image features. However, since there is a large difference in distribution between medical images and general images, fine-tuning is performed using VQA-Med2019 data [9]. We extract question features from pretrained BERT-Base, Cased [10]. All the questions are then embedded by the WordPiece which is used by BERT. On the other hand, all the answers are embedded by one-hot encoding. Proposed model consists of DNN, and detailed architecture of DNN is mentioned in subsection 4.3.

4.1 Image Pre-processing

We process image normalization, standardization and inpainting [11]. We show the flow of image pre-processing in Fig. 3. Firstly, all the images are grayscaling and resizing at 255×255 shape. Secondly, we make masks for inpainting in following four steps.

- Casting laplacian filter on resized images.
- Binarizing images with a threshold 50.
- Closing images with kernel size 5.
- Opening images with kernel size 3.

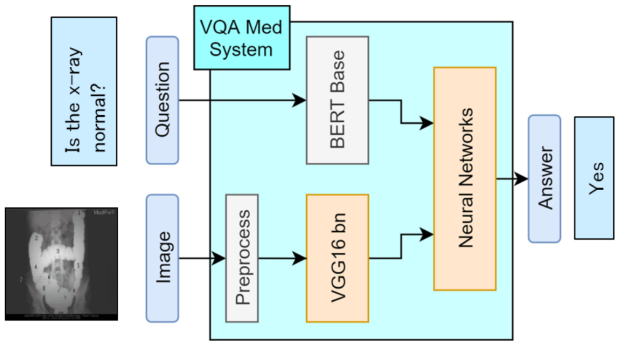

Fig. 2. Overview of our VQA-Med System

Thirdly, we cast inpainting images using the masks. We illustrate Fig. 4 where you can compare the raw images with the inpainting images. Finally, we make center crop images at 224×224 and normalize images as described in [8].

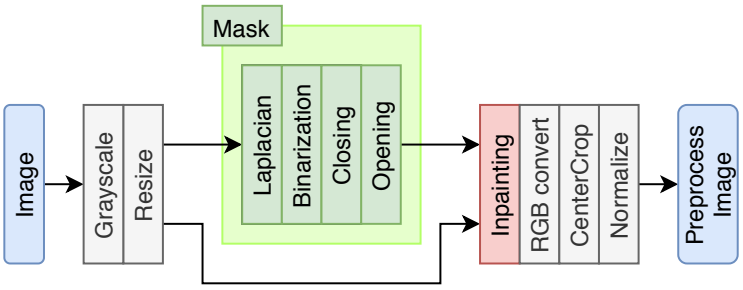

Fig. 3. Overview of our Image Pre-processes

4.2 Pre-training

Our networks, illustrated at left of the Fig. 5, classify the images of VQA-Med2019 into each attribute as a pre-training task. VQA-Med2019 dataset consists of question-answering classified into 4 categories of Modality, Plane, Organ and Abnormality per image. Similar to VQA-Med2020, question pattern is typical, and the answer can be predicted from the image alone, almost regardless of the question. We regard the answer of each category except Abnomality attached to the image as the attributes of the image, and perform the task of classifying each attribute of the image. However, for Modality, the classification

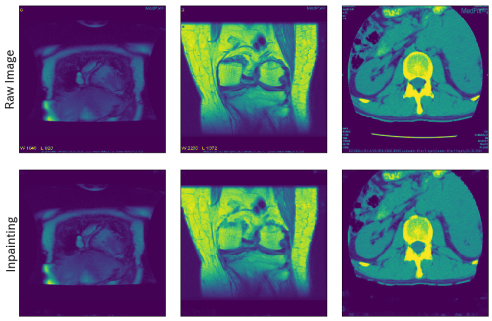

Fig. 4. Raw images and Inpainting images

is too subdivided, so use the rough classification given in the VQA-Med2019 dataset paper [9]. The pre-training model consists of VGG16 and two FC layers. We input the pre-processed image in Section 4.1 into VGG16 and obtain 4096-d features and multiply the matrix $W_1 \in \mathbb{R}^{4096\times1000}$ by the 4096-d features, and perform batchnorm [12], ReLU [13] and dropout [14] ratio= 0.5 at FC1, and obtain 1000-d features. Then we multiply the matrix $W_2 \in \mathbb{R}^{1000\times27}$ by the 1000-d features and obtain each attribute probability of softmax function. W_1, W_2 and VGG16 have trainable parameters.

4.3 Architecture of Proposed Model

Proposed model, illustrated right of the Fig. 5, generates an answer as a classification problem. Proposed model has VGG16 and FC1, 2, 3, 4, the weights of FC1, 2, VGG16 is trained by a pre-training task. VGG16 weights are frozen, and FC1, 2 are fine-tuned. Only FC3, 4 are trained from the beginning. FC3, 4 are based on the Gated Linear Unit (GLU) [15], and FC3 consists of $W_3 \in \mathbb{R}^{1000\times332}$, bias $b \in \mathbb{R}^{332}$. FC4 consists of matrix $W_4 \in \mathbb{R}^{795\times332}$, batchnorm and sigmoid. Then outputs of FC3, 4 are fused by element-wise multiplication and obtained using the softmax function to get the probability of answers.

GLU masks each dimension of the features obtained by FC3 with a real number from 0 to 1. The introduction of GLU is based on the consideration that it is possible to narrow down the answers to some extent only from the attribute information of question texts and images.

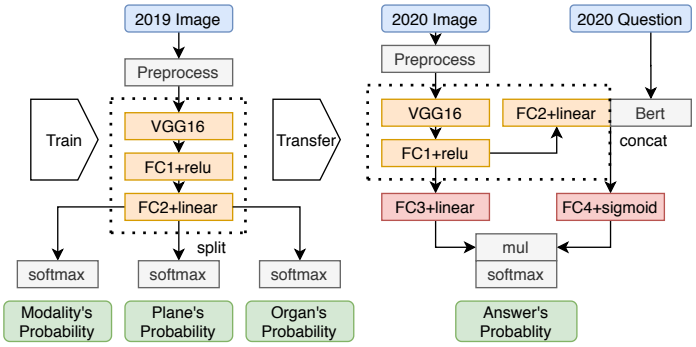

Fig. 5. Pretrained model and Proposed model

5 Experiments

Experiments are also performed on the baseline model, proposed model and proposed model without pre-training to show the usefulness of the proposed model. We first describe the baseline model in subsection 5.1, followed by the description of experimental conditions and evaluations in subsection 5.2. We finally describe experimental results and computational scores in subsection 5.3.

5.1 Baseline Model

The overview of the baseline model is shown in Fig. 6, feature fusion is used by concatenation instead of GLU. Baseline model has VGG16, FC1, 2 and FC3. FC3 has weights
$W \in \mathbb{R}^{1895\times 332}$. Other layers are in accordance with the proposed model in subsection 4.3.

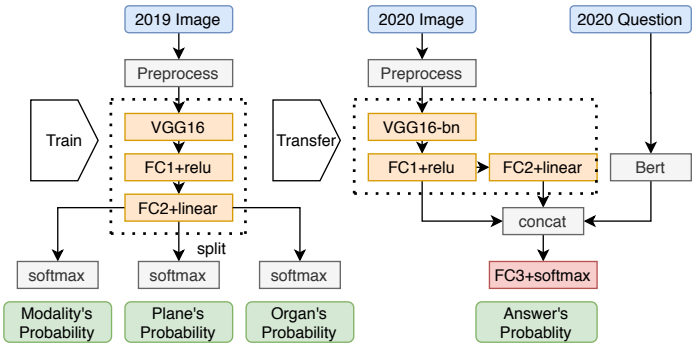

Fig. 6. Pretrained model and Baseline model

5.2 Conditions and Evaluations

We train our models using the train set and verify training model by the validation set. We determined the following hyper-parameters; loss function as cross entropy loss, the number of epoch 300, batch size of 64, optimizer as RMSprop [16] with a learning rate of 0.001. When training, we shuffle the training set order for each epoch, and training images are randomly flipped left and right with probability of 0.5.

The VQA-Med task adopts two evaluation method, accuracy and BLEU [17]. BLEU score measures the similarity between the predicted and correct answers.

5.3 Results

We submitted the baseline model and the proposed model and obtained the evaluation on the test set.

The results of our models show in Table 2. These results show that the fusion method using GLU is superior to the concatenation fusion, and according to the verification results, it can be seen that the accuracy is slightly improved by the pre-training task. VQA-Med2020 competition result is shown in Table 3, and our rank is 8th.

Table 2. Exprimental Results

Model	Val Accuracy	Test Accuracy	Test BLEU
Baseline	0.392	0.282	0.331
Proposed	0.412	0.314	0.350
Proposed -pre-training	0.408	-	-

Table 3. VQA-Med2020 Competition Results

Rank	Participants	Accuracy	BLEU
1	z liao	0.496	0.542
2	TheInceptionTeam	0.480	0.511
3	bumjun jung	0.466	0.502
4	going	0.426	0.462
5	NLM	0.400	0.441
6	harendrakv	0.378	0.439
7	Shengyan	0.376	0.412
8	**kdevqa**	**0.314**	**0.350**
9	sheerin	0.282	0.330
10	umassmednlp	0.220	0.340
11	dhruv sharma	0.142	0.177

6 Conclusion

In this research, we describe the models we submitted in ImageCLEF2020 VQA-Med task. We proposed a model of feature connection by GLU and a pre-training task by VQA-Med2019 dataset. We also introduced the removal of a logo and texts using inpainting as image pre-processing. We show that fusion of functions using GLU is superior to simple concatenation, and slightly improved score using pre-training task. Proposed model scores 0.314 in accuracy and 0.350 in BLEU in VQA-Med2020 task, and our rank is 8th.

Acknowledgment

A part of this research was carried out with the support of the Grant-in-Aid for Scientific Research (B) (issue number 17H01746).

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. 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.
3. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.
4. 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. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
5. Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.
6. Ross Girshick. Fast r-cnn. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 1440–1448, Washington, DC, USA, 2015. IEEE Computer Society.

7. Xin Yan, Lin Li, Chulin Xie, Jun Xiao, and Lin Gu. Zhejiang university at imageclef 2019 visual question answering in the medical domain. In *CLEF*, 2019.
8. Torchvision.models. https://pytorch.org/docs/master/torchvision/models.html.
9. 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 *CLEF2019 Working Notes*, CEUR Workshop Proceedings, Lugano, Switzerland, September 09-12 2019. CEUR-WS.org <http://ceur-ws.org>.
10. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. cite arxiv:1810.04805Comment: 13 pages.
11. Alexandru Telea. An image inpainting technique based on the fast marching method. *Journal of Graphics Tools*, 9(1):23–34, 2004.
12. Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 448–456. JMLR.org, 2015.
13. Abien Fred Agarap. Deep learning using rectified linear units (relu), 2018. cite arxiv:1803.08375Comment: 7 pages, 11 figures, 9 tables.
14. Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, January 2014.
15. Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 933–941. JMLR.org, 2017.
16. T. Tieleman and G. Hinton. Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, 2012.
17. 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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.