IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 14, 2020, accepted March 6, 2020, date of publication March 11, 2020, date of current version March 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2980024

CGMVQA: A New Classification and Generative Model for Medical Visual Question Answering

FUJI REN, (Senior Member, IEEE), AND YANGYANG ZHOU
Faculty of Engineer, University of Tokushima, Tokushima 770-8506, Japan
Corresponding author: Fuji Ren (ren@is.tokushima-u.ac.jp)

ABSTRACT Medical images are playing an important role in the medical domain. A mature medical visual question answering system can aid diagnosis, but there is no satisfactory method to solve this comprehensive problem so far. Considering that there are many different types of questions, we propose a model called CGMVQA, including classification and answer generation capabilities to turn this complex problem into multiple simple problems in this paper. We adopt data augmentation on images and tokenization on texts. We use pre-trained ResNet152 to extract image features and add three kinds of embeddings together to deal with texts. We reduce the parameters of the multi-head self-attention transformer to cut the computational cost down. We adjust the masking and output layers to change the functions of the model. This model establishes new state-of-the-art results: 0.640 of classification accuracy, 0.659 of word matching and 0.678 of semantic similarity in ImageCLEF 2019 VQA-Med data set. It suggests that the CGMVQA is effective in medical visual question answering and can better assist doctors in clinical analysis and diagnosis.

INDEX TERMS Classification model, generative model, medical image, transformer, visual question answering.

I. INTRODUCTION

Health has consistently been one of our most concerned issues. So far, there are still few ways to easily learn about our physical conditions without professional guidance. Medical imaging, as a non-invasive technique for producing images of the internal aspects of the body, is an extremely important tool for doctors to understand our physical conditions in clinical analysis and diagnosis. However, the information obtained from medical images by different doctors can differ. Deep learning, as a powerful information processing tool, plays an increasingly significant role in health informatics [1]. Some deep learning methods can be used to extract information from images or texts. Specific to the medical field, a good medical visual question answering (VQA) model based on deep learning can automatically extract the information contained in the medical images and assist in medical diagnosis. Meanwhile, it can help patients to get a preliminary understanding of their physical condition through the medical VQA model, which can devote in choosing a more targeted medical treatment plan. In general, using the medical VQA model can alleviate the problems caused by the imbalanced distribution of medical resources.

In recent years, the application of computer-aided diagnosis (CAD) methods for processing medical information are becoming widespread [2]. However, lots of the CADs concentrate on lesion diagnosis or segmentation of a single type of medical images, such as tumor tracking [3]. Some of the CADs focus on medical records to predict risks [4]. Most CAD methods are aimed at the diagnosis of a single disease, including lung disease [5], breast cancer [6], etc.. However, there is less work to combine natural language processing with medical image processing, such as medical image caption [7] and medical VQA [8]. There are few public medical VQA data sets. A typical one is VQA-RAD [9], including 315 medical images from MedPix®, https://medpix.nlm.nih.gov/. We have applied for permission to use another data set called ImageCLEF 2019 VQA-Med [10]. The data from this set contain a series of questions from the elementary to the profound. Elementary questions refer to the ones that patients or medical students may ask, such as the type of the scan. Profound questions involve the severity of illness or diagnosis. This data set involves a wide range of medical images and question-answer pairs, close to the real medical environment. The ImageCLEF

The associate editor coordinating the review of this manuscript and approving it for publication was Victor Sanchez.

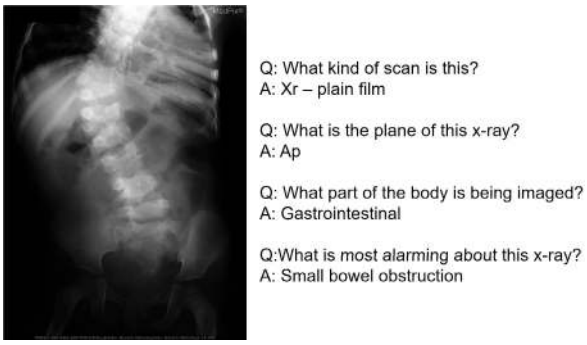

FIGURE 1. An example in the ImageCLEF 2019 VQA-Med (a medical image with several questions and answers).

2019 VQA-Med has 10 times more images than the VQA-RAD. As the example shown in FIGURE 1, there are a variety of questions that can be considered in a medical image. There are various types of medical images, like magnetic resonance imaging, computerized tomography, etc., and each image can be taken from a different angle. The accuracy of medical VQA so far is lower than the level of human doctors, owing to the variety of answer expression and the difficulty of answer evaluation. There is still no existing method that can solve this comprehensive problem well.

VQA is the task of answering relevant questions based on the contents of images, involving image processing and natural language processing techniques. The process of the VQA task can be divided into three parts: extracting image features, extracting question features, and integrating features. Generally, the method of extracting image features is to use transfer learning [11] as a feature extractor to deal with images such as deep residual network [12] (ResNet). Extracting question features mainly uses the recurrent neural networks [13] like long short-term memory [14] to turn texts into vectors. Features integration can be used in two kinds of approaches: classification and generation. Integration methods for classification can be based on concatenation [15], bilinear pooling [16] or some other mechanisms. Integration methods for generation are commonly based on encoder-decoder [17] framework.

Answering the questions in VQA requires an understanding of vision, language and common sense knowledge. Background knowledge can supplement relevant information in a certain field [18]. To promote understanding, some models employ knowledge-based reasoning for VQA [19]. Existing VQA models perform well when there is no need for them to fully understand the image information, such as object detection. However, they cannot solve different kinds of complex questions at the same time. Existing CAD methods cannot achieve satisfactory results when dealing with different types of images like magnetic resonance and computerized tomography, as well as different organs like brain, lung, and so on at the same time. This is similar to the challenge in the VQA task we described.

In order to meet this challenge, we divide the complex medical VQA task into multiple simple tasks, and introduce a model that can do both classification and answer generation. The core of our model is the multi-head self-attention mechanism [20], which can learn the internal representations from the training data. This mechanism uses the multi-channel parallel computation and carries out the linear transformation and the weight calculation on each term of the input itself. We reduce the parameters based on the transformer and introduce the weight sharing between embedding and output layers to do lower memory consumption training.

Bidirectional Encoder Representation from Transformers [21] (BERT) is a pre-trained language representation model, coping with mask language model tasks and next sentence prediction tasks at the same time. Inspired by that, we adjust the masking and output layer to unify classification and answer generation into one model, as shown in FIGURE 2. Our proposed model can do classification by using only images and questions, and do generation by using images, questions and unidirectional masked answers. We adopt the convolution outputs of different layers from ResNet152 as the image input features. Text input features we used is the word piece [22] features after three kinds of embeddings.

CGMVQA is used to answer questions containing various medical images by transforming the strong artificial intelligence problem into multiple weak artificial intelligence problems. The data set is ImageCLEF 2019 VQA-Med. We use accuracy to evaluate the predicted answers and get a score of 0.640, which is better than that of the task challenge winner (0.624 in TABLE 2). We also employ Bilingual Evaluation Understudy [23] (BLEU) and Word-Based Semantic Similarity [24] (WBSS) to evaluate the predicted answers and get results of 0.659 and 0.678 respectively. These results indicate the effectiveness of our model.

Our contributions can be summarized as follows:

- We propose a model for medical VQA. This model can switch between the classification model and the generative model by modifying the output layer and the loss function without modifying the core part. Our model establishes new state-of-the-art results on ImageCLEF 2019 VQA-Med data set.
- We adopt the pre-trained ResNet152 to extract image features. Because the structure of the medical image is relatively fixed, we extract features from different convolutional layers so as to retain the image semantic information from different dimensions, which is different from other models.
- Unlike other models, we abandon the traditional encoder-decoder framework, and build the generative model through the method of masking position by position.
- Compared to BERT-like models, we reduce the parameters of our model by weight sharing and embedding factorization. Our model can be implemented on a single GPU.

The rest of this paper is structured in the following part. Section II briefly reviews the related work.

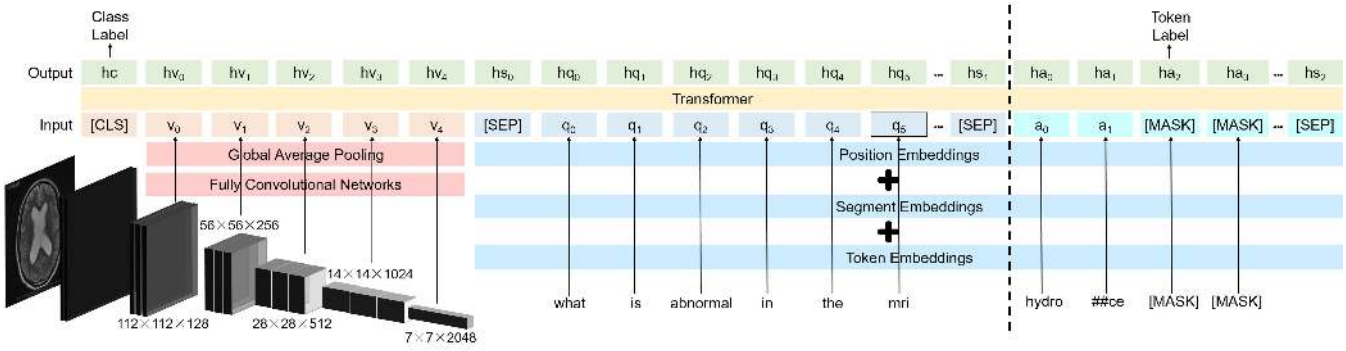

FIGURE 2. Overview of CGMVQA architecture, including image embedding (red), text embedding (blue), model core (yellow) and output (green).

Section III provides the details of CGMVQA. Section IV analyzes the experiments and the results of our model. Section V presents the discussion, followed by the conclusion in section VI.

II. RELATED WORK

Initialization is important for model training. In computer vision, transfer learning based on ImageNet [25] is often used as a feature extractor in other computer vision tasks. The pre-trained model without the classification layer has a powerful representation ability. ResNet152 is one of the best performing neural networks in ImageNet classification tasks. Considering that the distribution gap between ImageNet and medical images is too large, we do not directly use the transfer learning model as a feature extractor. Besides, U-Net [26] with few parameters and skip connection, is appropriate for medical image semantic segmentation. In U-Net, the output of each convolutional layer is the feature of different dimensions of the image. We learn the idea from U-Net and extract different convolutional layer outputs from the pre-trained model Resnet152. Then we use these features as the input of the image part, as shown in FIGURE 2.

Word embedding can represent words as vectors for a better training. Global vectors [27] is an unsupervised learning algorithm, which maps words into meaningful space based on semantic similarities and uses vector representation. Since the vector of each word is fixed, this method cannot disambiguate according to the context. Embedding from language models [28] is a way to learn the context of words by deep networks. Compared with global vectors, this method adjusts the word vector according to the context, but it requires a high computational cost. ALBERT [29] proposed factorized embedding parameterization, which can greatly reduce the number of embedded parameters. We adopt this technique and share the weight of output embeddings with the input embeddings when generating the answers.

In natural language processing, pre-trained models like GPT [30] and BERT have achieved lots of breakthrough results. These models all use the multi-head self-attention mechanism in transformer, which is good at learning contextualized text representations. Existing research [31] shows that placing the layer normalization of the transformer into the residual part can avoid adjusting the learning rate through the warm-up optimizer. We try to combine the features of images and texts through this mechanism.

Recently, multi-modal pre-trained models like ViLBERT [32] and LXMERT [33] have adopted two transformers to deal with images and texts independently. Others like VL-BERT [34] and VisualBERT [35] have used one transformer to model visual-language representation. The key difference between CGMVQA and the other models is that the other models can only do classification, while our model can do both classification and answer generation by modifying the output layer. In addition, considering the closed-domain data set and training cost, we do not use the large-scale pre-training process, but directly used the model for training. And our parameters are much less than those of the above models.

Most existing VQA data sets are open-domain, such as VQA challenge data set [36]. Some answers in this data set only require common sense knowledge, instead of the image itself. For instance, the question ''What color is the tree?'' appears 73 times in the data set, and 95.9% of the answers are ''green''. Unlike the open-domain sets, the medical data sets are closed-domain, and have a smaller amount of data. Most of the public medical data sets only contain a single disease, with labels like ''sick or not'' or ''severity of illness'' [37]. ImageCLEF 2019 VQA-Med data set involves various image types and organs, which is more complicated.

Different VQA data sets have their own evaluation metrics, but these metrics are mainly variants of accuracy [38] [39]. Both accuracy and BLEU are used in ImageCLEF 2019 VQA-Med tasks. We also adopt WBSS [24] to compare the degree of semantic similarity. In addition, we use the metrics such as precision and recall for the classification part.

III. CGMVQA

The problem can be formulated as follows:

Given an image V_i, and a question $Q_i = (q_{i0}, q_{i1}, \cdots, q_{ik})$, the goal is to get an answer A_i. For the classification problem, the answer A_i has candidate items, which are denoted as $A_i \in \{A_0, A_1, \cdots, A_m\}$; for the generative problem, there is no candidate items but a word pieces vocabulary W for

Algorithm 1: CGMVQA Pseudo Algorithm

Input: image V_i, question Q_i, max length L, learning rate η;
Output: classification answer A_i or generative answer $a_0, \cdots, a_{L-1}$;

Initialize the parameters θ of the model f randomly;
$v_0, \cdots, v_4 \leftarrow V_i$ features from pre-trained ResNet152;
$q_0, \cdots, q_{L-1} \leftarrow Q_i$ features from Embeddings;
if *Classification mode* **then**
    Cross-entropy loss function $Loss_\theta \leftarrow \ell_{(\theta)} = -\sum_{j=0}^{m} A_j \log f(v_0, \cdots, v_4, q_0, \cdots, q_{L-1}; \theta)$;
    (Train θ until it converges.) Backpropagation $\theta \leftarrow \theta - \eta \frac{\partial Loss_\theta}{\partial \theta}$;
    $A_i \leftarrow$ Class label $h_c(\theta)$;
    Inference: return A_i;
else if *Generative mode* **then**
    $a_0, \cdots, a_{L-1} \leftarrow$ [MASK] features from Embeddings;
    for $j = 0; j < L; j++$ **do**
        $Loss_\theta \leftarrow \ell_{(\theta)} = -\sum_{a_j \in W} a_j \log f(v_0, \cdots, v_4, q_0, \cdots, q_{L-1} a_0, \cdots, a_{L-1}; \theta)$;
        (Train θ until it converges.) $\theta \leftarrow \theta - \eta \frac{\partial Loss_\theta}{\partial \theta}$;
        $a_j \leftarrow$ Token label $h_{aj}(\theta)$;
    Inference: return $a_0, \cdots, a_{L-1}$;

answer A_i, and we denote the answer $A_i = (a_{i0}, a_{i1}, \cdots, a_{in})$, $a_{i0}, a_{i1}, \cdots, a_{in} \in W$.

We propose a model called CGMVQA for medical VQA. As is illustrated in FIGURE 2, we use this model to combine images and texts. Depending on the different output layers, the model can be constructed as a classifier or generator for downstream tasks. The classification mode uses images and questions (the model only contains the left part of the dotted line). The class is predicted by the first item of output h_c. The generative mode includes images, questions, and masked answers. The next word piece is predicted by the output of the first mask label h_{aj}. On the ImageCLEF 2019 VQA-Med data set, this mode is used to answer the questions of yes-no, modality, plane and organ system. And the generative mode is used to give the answers of abnormality questions, since there is no candidate answer in this category.

A. BASELINE

We had a submission in the ImageCLEF 2019 VQA-Med task, and got ranking fourth [40]. We choose that model as one of the baseline. In the experimental part, the CGMVQA not only exceeds our previous work, but also achieves new state-of-the-art results. We use a simple classifier to divide all the questions into four different categories in the baseline. InceptionResNetV2 [41] uses the technology of inception and residual connection to expand the width and depth of the convolutional neural network. We adopt the pre-trained InceptionResNetV2 as the image feature extractor, and the pre-trained BERT as the text feature extractor in this baseline. We concatenate these features and train them by a multi-layer perceptron for classification. As for generative part, we concatenate these features as the initial state of the decoder, and train them by a long short-term memory.

The other baseline is the Bottom-Up and Top-Down Attention model [42]. This model has obtained state-of-the-art results in some open-domain tasks such as the VQA Challenge [36]. According to the paper, this model concatenates the image features and the embedding of the question and uses them to generate the top-down attention weight. In order to use this model for the ImageCLEF 2019 VQA-Med data set, we do not use the pre-trained image features provided by the author, but keep the same with our proposed model (using the image features that we extract by ResNet152). In order to make the comparison experiments with similar numbers of parameters, we reduce the hidden size of this model proportionally. The output is generated by a multi-label classifier, which means that the model can only be used for classification.

B. PRE-PROCESSING

Considering the wide distribution of training data and training difficulties caused by plenty of categories, we try to simplify this complex problem. Since the category is not given in the test set, we use a simple classifier to divide the questions into five different categories: yes-no, modality, plane, organ system and abnormality, similar to what we used before [40]. Unlike baseline, yes-no questions are trained separately. The simple classifier is sufficient to correctly classify the data only through features of the questions. Different categories of questions have their own characteristics, for example, yes-no questions begin with words including ''does'', ''is'', etc..

Modality can be regarded as a fine-grained classification task. Its candidate answers include more than mri and ct, as shown in FIGURE 5. Plane and organ system categories are simple to do classification due to their few kinds of candidate answers and the similar composition of images with the same

candidate answer. Abnormality involves more complicated issues such as the location of the lesion. Most questions have their own answers, so this category is hard to do classification. We try to deal with this category in answer generation.

We adjust the image to the same size: 224 × 224×3. For classifications, data imbalance can make model training more difficult. We extend the images of the candidate answers in each category to the same number by data augmentation [43]. Data augmentation can also increase the generalization ability of the model. Specifically, data augmentation includes random rotations within ±10°, small amplitude random shearing, scaling, and horizontal, vertical shifting. We do not use flipping because there are lesion location questions in the data set.

As for the texts, we convert all the questions and answers into lower case letters and remove the punctuation. Since word-based token will cause a large dictionary, we use the word piece tokenization method, like BERT model.

C. EMBEDDINGS

Different convolutional layers have different feature extraction capabilities. ResNet152 has five blocks that resize the image through convolution kernels. We attempt to extract the features of the images from the five convolutional layers in ResNet152. The feature sizes after extraction are shown in FIGURE 2. We adopt untrained Full Convolutional Networks [44] to unify the number of feature maps and use the Global Average Pooling [45] strategy to unify the dimensions.

We add three kinds of embeddings to express different word pieces. We employ a decomposition method in the token embeddings part. Specifically, we project the word pieces into a lower dimensional space (here we use size 128), then project them to the hidden space. In the CGMVQA, this method reduces the embeddings parameters by 60% compared to direct projection to the hidden space, while the accuracy only decreased by 0.008. Segment embeddings are used to distinguish between the questions and masked answers. When models are employed to classify, only questions are involved in training. Position embeddings are used to represent the order of each word piece in the sequence.

We separate the embedding images, questions and answers with ''[SEP]'' token and add a ''[CLS]'' token for classification at the beginning.

D. TRANSFORMER

The transformer has the advantage of parallel computing and is gradually replacing LSTM to deal with sequence problems. As shown in FIGURE 3, we make some improvements based on the original transformer. Similar to BERT, we use the Gaussian Error Linear Unit [46] activation function in the fully-connected feed-forward network. To avoid using the warm-up optimizer, we put layer normalization before the multi-head attention layer and the fully-connected feed-forward network. We adopt a residual connection here to avoid the vanishing gradient problem. Inside the dotted box is a transformer block. To further decrease the number

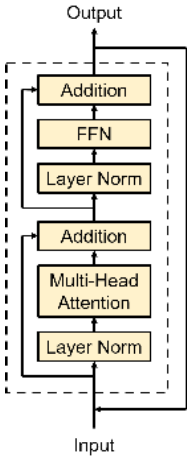

FIGURE 3. Architecture of pre-layer-normalization multi-head self-attention and feed-forward network transformer.

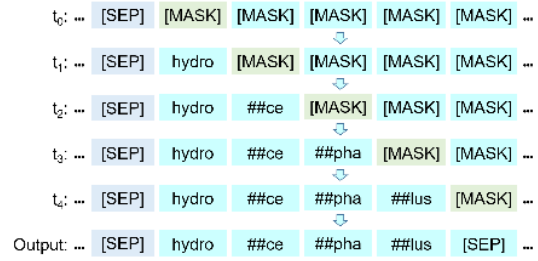

FIGURE 4. The prediction process of generative model. Each line is the output of the previous time and the masked input of the current time, and the green box is the position of the predicted word piece.

of parameters, we try to share the weight among the blocks, but the accuracy drops significantly.

E. CLASSIFICATION MODE

According to the process we described above, we train the images and corresponding questions, and use the special token ''[CLS]'' to do classification. We add fully-connected layers with the tanh activation function on the top of our model to calculate the possibility that the ''[CLS]'' belongs to a candidate answer. The estimated probability of the model is:

$$P_{(A=A_i|V_iQ_i;\theta)} = \frac{e^{\theta_i^T V_i Q_i}}{\sum_{j=1}^{m} e^{\theta_j^T V_i Q_i}} \quad (1)$$

where V_iQ_i is the ith of image and question features, A_i is the corresponding answer, and θ_j is the weight vector of the jth class. We use the classifiers to target different categories.

F. GENERATIVE MODE

There is no candidate answer to the abnormality question, which is different from that of other categories. We employ the generative mode to obtain the answer. Unlike the classification mode, we add the masked answer in generative mode training. When predicting a word piece, the current word piece and the following pieces are masked to avoid information leak. For example (FIGURE 4), besides the features of

the image and the question, we use the features of all masked tokens to predict the first word piece h_{a0} (''hydro''), then use ''hydro'' and other following masked tokens to predict h_{a1}. Loop this process until the special token ''[SEP]'' is obtained.

We use fully-connected layers with the Gaussian Error Linear Unit activation function on the top of our model to predict the word piece at the current position. Besides, we use beam search [47] to generate the answer when predicting, with beam width in 5. Beam search is a greedy algorithm that explores the best combination by extending the most promising nodes in a limited set [48]. In the prediction process, we add a penalty factor in the prediction process, to make the generated results look more readable:

- When the result continuously generates the same word, such as ''$\cdots$ of of the knee'', we reduce the probability of the extra duplicate words (''of'').
- Because of using word piece tokens, we reduce the probability of the suffix with ## from the beginning of the generated results, like ''##ce''.

G. MODEL SETUP

In order to run the CGMVQA on a single GPU, we try to reduce the number of parameters in our model. To balance efficiency and information loss, we set the maximum input length of the questions and answers to 12. We set the hidden size to 312 and embedding size to 128. We share the weight between the token embeddings and the output layer in generative mode. For the transformer part, we set 12 heads in the multi-head self-attention mechanism. We put the layer normalization before the self-attention layer and the feed-forward layer, with $L2 = 1 \times 10^{-12}$. And we set 4 transformer blocks. TABLE 1 shows the process of setting the hyper-parameters.

As for training, we set the learning rate to 0.0001 in ADAM [49] optimizer, with no dropout and $batchsize = 64$. To avoid gradient exploding problem, we use gradient clipping in training. The total parameters of our model are 6.4M, which is greatly reduced compared to 108M of the original BERT, which can be easily implemented on a GPU.

TABLE 1. Hyper-parameters setting experiment.

	Accuracy
Reference model	0.640
Reference: hidden size = 312	
Hidden size = 192	-0.008
Hidden size = 432	**+0.002**
Reference: embedding size = 128	
Embedding size = 64	-0.010
Embedding size = hidden size	**+0.008**
Reference: heads = 12	
Heads = 6	-0.016
Reference: blocks = 4	
Blocks = 2	-0.018
Blocks = 6	-0.004
Reference: no parameter sharing	
Attention layers sharing	-0.006
Feed-forward layers sharing	-0.012
blocks sharing (both attention and feed-forward)	-0.022
Reference: pre layer normalization	
Post layer normalization	-0.008
Reference: no dropout	
dropout = 0.1	-0.010
Reference: $L2 = 1 \times 10^{-12}$	
$L2 = 1 \times 10^{-11}$	-0.014
$L2 = 1 \times 10^{-13}$	-0.002

IV. EXPERIMENT AND EVALUATION

A. DATA ANALYSIS

Compared with open-domain images, medical images are not only more fixed in structure, but also have a smaller amount of data. There are few data sets of VQA in the medicine domain. We use the ImageCLEF 2019 VQA-Med data set here. This data set has 12792 pairs of question-answer and 3200 medical images for training; 2000 pairs of question-answer and 500 images for validation; 500 questions and 500 images for testing. In the training set and validation set, each image may correspond to 4 different categories of questions: modality, plane, organ system and abnormality. Each image in the test set corresponds to only 1 question, with no category marked. Candidate answers are provided in the first three categories of questions: 43 kinds of answers in modality category, 16 kinds of answers in plane category and 10 kinds of answers in organ category. Owing to the various types of diseases, abnormality questions do not have candidate answers.

During the analysis of the data, we find that there are a few yes-no questions in modality and abnormality categories. We take them out and put them into a new category. We also note that there are some wrong labels in the data set, and we do not make any change, just regard them as noise. The candidate answers cannot be in one-to-one correspondence with every ground truth in the data set. Some candidate answers even have a similar meaning, like ''ct with iv contrast'', ''ct w/contrast (iv)'' and ''iv''. During training, we do not deal with them specifically. In the result visualization part, we merge them into one item. Furthermore, the class imbalance problem can be seen from the data set. Especially in the modality category (FIGURE 5 on the next page), about half of the classes are not included in the test set.

B. EVALUATION METRICS

Accuracy is a commonly used evaluation metric in VQA. We adopt a strict accuracy to compare the difference between the output results and the ground truths. For the classification method, we use the precision, recall, F1 score from scikit-learn [50] and confusion matrix to help analyzing the results. Since we employ the generative method in the abnormality category, the accuracy cannot reflect the effect of the model well, so we introduce other metrics as follows.

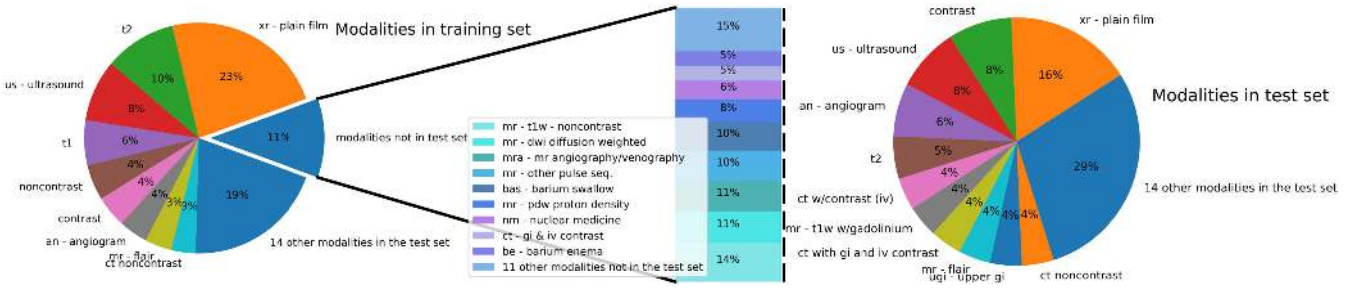

FIGURE 5. The pie of modality category data distribution.

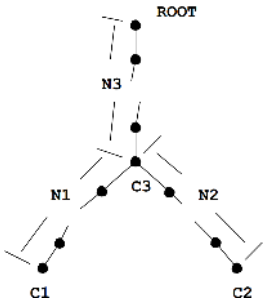

FIGURE 6. The concept similarity measure from references [51].

WBSS is an algorithm for calculating semantic similarity in the biomedical field based on Wu-Palmer similarity [51].

$$Sim_{WP}(C1, C2) = \frac{2 * N3}{N1 + N2 + 2 * N3} \tag{2}$$

According to the paper [51] and FIGURE 6, $C3$ is the least common super concept of $C1$ and $C2$. $N1$ is the number of nodes on the path from $C1$ to $C3$. $N2$ is the number of nodes on the path from $C2$ to $C3$. $N3$ is the number of nodes on the path from $C3$ to root. For the abnormality category, the higher the WBSS score is, the higher the similarity between the semantics of the generated answer and the ground truth would be.

Bleu is another automatic evaluation metric to compare the frequency of co-occurrence words, often used for machine translation. Unlike WBSS, BLEU does not concern similar expressions, only concerns word matching.

$$BLEU = BP \cdot exp(\sum_{n=1}^{N} w_n \log p_n) \tag{3}$$

According to the paper [23], BP is the brevity penalty, p_n is the geometric average of the modified n-gram precision. N-grams here is up to length $N = 4$ and positive weight $w_n = 1/N$. We use the natural language toolkit [53] to calculate the BLEU score. However, the score may be unstable once there's no 4-gram match between the reference and the hypothesis.

C. RESULTS

We compare the CGMVQA with other methods in TABLE 2. We report averages by training each method with 3 different random seeds. In 500 outputs of the test set, there are no more than 10 different predicted answers each time (± 0.01 in strict accuracy). Our model achieves 0.640 accuracy score, 0.659 BLEU score and 0.678 WBSS score. Good option refers to the option that appears most frequently in each category of the training set as the answer of the test set (all "no" in yes-no category; all "xr - plain film" in modality category; all "axial" in plane category; all "skull and contents" in organ system category; all "meningioma" in abnormality category). As can be seen from the good option, there is a severe data imbalance in the test set, especially in the plane category. Our model has a huge improvement than the good option.

Since there are few models that can do both classification and answer generation, we use the previous model that we proposed in the ImageCLEF 2019 VQA-Med task as the baseline [40]. Compared to this baseline, our model increase accuracy by 3.4%, BLEU by 2.5%, and WBSS by 3.1%. Because the Bottom-Up and Top-Down Attention model can only be used for classification, we use this baseline to answer the questions of yes-no, modality, plane and organ system. The results of this model in both modality and plane categories exceed those of our previous model, but do not exceed those of the CGMVQA. In the ImageCLEF 2019 VQA-Med task, the best method [52] achieved 0.624 accuracy score and 0.644 BLEU score (no WBSS score). The CGMVQA also gets better results than that method.

In the ablation experiments, we add the model without pre-classification (no pre-class) and the model without data augmentation (no data-balance) to the comparison. No pre-class means that the model does classification directly from all the candidate answers when being trained (including abnormality category). The results of models without pre-classification have a certain degree of decline in every category. Compared with the no data-balance model, our proposed model improves the result of the data imbalanced categories. However, there is no significant improvement in other categories (yes-no and abnormality).

We use additional evaluation metrics to evaluate the effect of the classifiers. The yes-no classifier achieves 0.781 in precision, recall and f1 score. The modality classifier with 44 candidate answers achieves 0.735 precision score, 0.679 recall score and 0.683 f1 score. We merge similar candidate answers (like "mr - flair" and "flair") in modality category and draw the confusion matrix, as shown in FIGURE 7.

TABLE 2. The performances in the comparative and ablation experiments.

Method	Yes-no			Modality			Plane			Organ			Abnormality			All		
	Accu	BLEU	WBSS	Accu	BLEU	WBSS	Accu	BLEU	WBSS	Accu	BLEU	WBSS	Accu	BLEU	WBSS	Accu	BLEU	WBSS
Good option	0.500	0.500	0.524	0.167	0.167	0.299	0.512	0.512	0.512	0.432	0.493	0.541	0.018	0.018	0.099	0.328	0.343	0.396
No pre-class	0.703	0.703	0.716	0.653	0.750	0.764	0.760	0.767	0.768	0.744	0.764	0.791	0.000	0.000	0.022	0.560	0.581	0.596
No data-balance	**0.781**	**0.781**	**0.792**	0.694	0.769	0.788	0.776	0.783	0.785	0.752	0.765	0.783	**0.044**	0.076	**0.124**	0.592	0.615	0.635
Previous work [40]	**0.781**	**0.781**	**0.792**	0.667	0.759	0.788	0.716	0.816	0.817	**0.784**	0.791	0.812	0.035	**0.088**	0.111	0.606	0.634	0.647
Up-Down [42]	0.719	0.719	0.732	0.806	0.871	0.873	0.824	0.834	0.835	0.744	0.751	0.772	-	-	-	-	-	-
Best in challenge [52]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.624	0.644	-
Ours	**0.781**	**0.781**	**0.792**	**0.819**	**0.880**	**0.886**	**0.864**	**0.864**	**0.866**	**0.784**	**0.797**	**0.819**	**0.044**	0.076	**0.124**	**0.640**	**0.659**	**0.678**

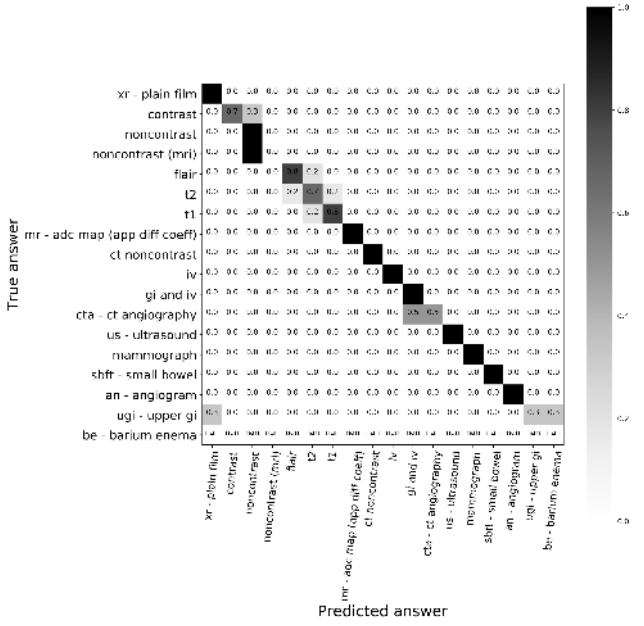

FIGURE 7. The confusion matrix of the modality category.

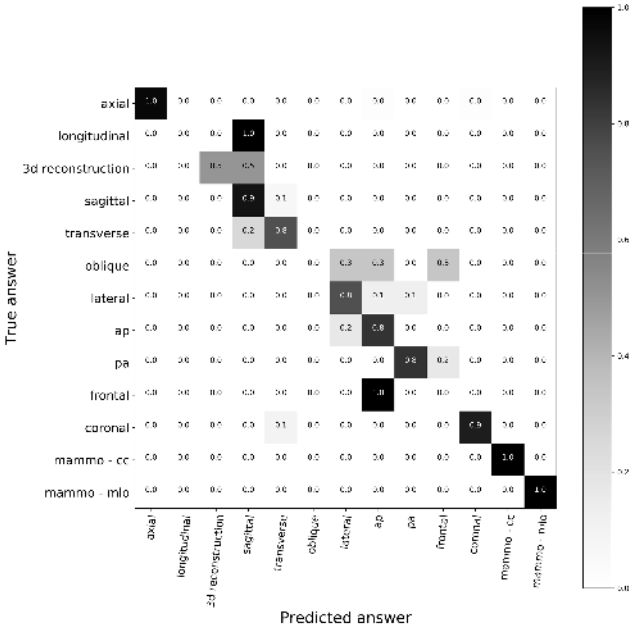

FIGURE 8. The confusion matrix of the plane category.

Most of the candidate answers can be successfully classified in this category. The model tends to predict noncontrast(mri) as noncontrast. These two concepts are overlapped. ''ugi-upper gi'' in the training set accounts for only 0.4%, which is quite different from the feature distribution in the test set. The classification accuracy of this class is only 30%. There is also a ''be-barium enema'' that is not successfully predicted.

The plane classifier with 16 candidate answers achieves 0.643 precision score, 0.651 recall score and 0.636 f1 score. We draw the confusion matrix, as shown in FIGURE 8. Some candidate answers are not included in the test set. In this category, both ''axial'' and ''mammo'' type can be accurately predicted. However, none of ''longitudinal'', ''oblique'' and ''frontal'' are accurately predicted. In contrast, the model is too confident to predict ''sagittal'' or ''ap''.

The organ system classifier with 10 candidate answers achieves 0.618 precision score, 0.647 recall score and 0.622 f1 score. We draw the confusion matrix, as shown in FIGURE 9. This classifier has the fewest candidate answers, but it does not perform better than other classifiers. The ''vascular and lymphatic'' with the least proportion in the training set, performs the worst in the test. The ''skull and contents'' with the most proportion in the training set, are too confident to be predicted.

In the training set, there are more than 1600 different answers to the 3000 questions in the abnormality category. We attempt to employ the classification mode training and get the same answer to all the questions on the test set. It means that the result will not be better than the good option. However, the amount of data (4k in total) is too small for a generative model (100k is reasonable in natural language processing). The generator we used achieves better accuracy than others, but 0.044 is not a high score. Additionally, due to the existence of synonym, strict accuracy cannot fully reflect the effect of the model. The ground truths of abnormality category are mostly composed of the words less than 4, so 4-gram match BLEU also cannot fully reflect the effect of the model.

V. DISCUSSION

The CGMVQA is efficient on the ImageCLEF 2019 VQA-Med data set. The images of this data set cover almost all human organ systems, as well as all categories of medical imaging. The questions of this data set cover the general

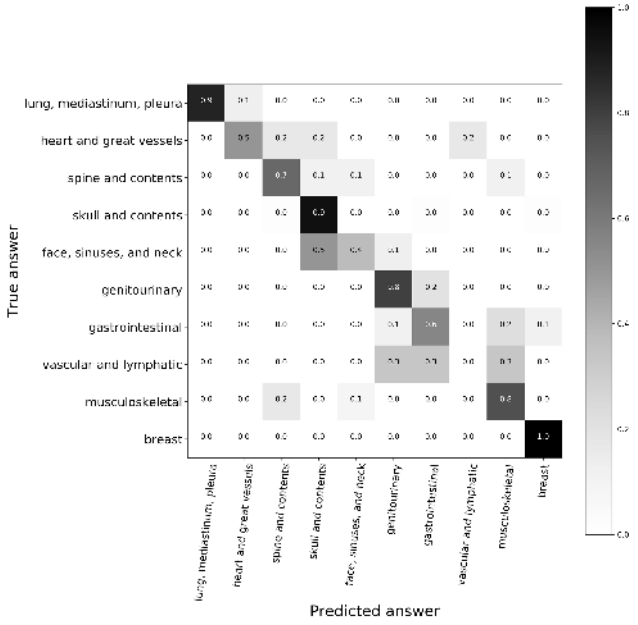

FIGURE 9. The confusion matrix of the organ system category.

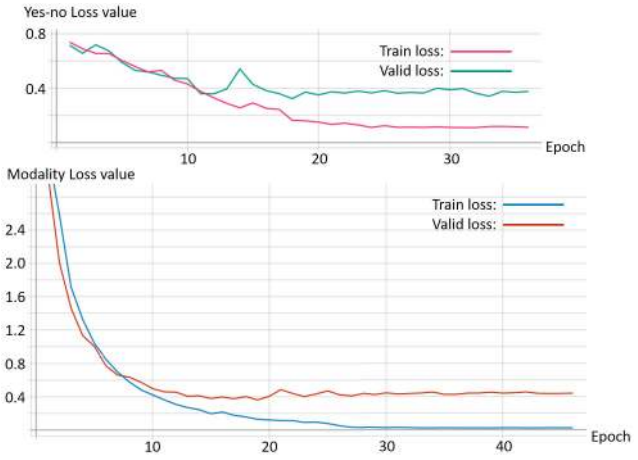

FIGURE 10. The loss curves of yes-no and modality categories.

questions that doctors need to know when reading medical images. Due to the lack of public VQA data in the medical domain, we do not experiment on other data sets.

Most of the deep learning models can only do classification or answer generation. We propose the method to divide the complex problem into multiple simple ones. The CGMVQA only needs to modify the input masking and output layer to do both classification and generation. Our model can achieve better results compared to the existing technology.

In the categories with a small number of classes (yes-no and organ system), our model is not significantly improved compared to the baseline. As can be seen from FIGURE 10, the loss value of the yes-no on the validation set is low, but the performance on the test set is not ideal. Because there is a problem that the data distributions in the test set and the training set are inconsistent, which often exists in data sets with a small amount of data. This problem leads to a bad

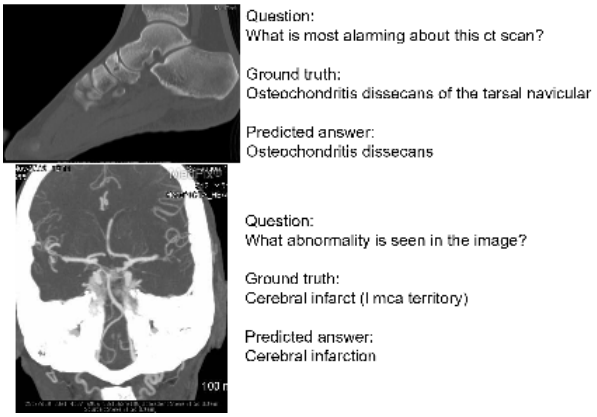

FIGURE 11. The examples of predicted answers in the abnormality category.

performance when testing in spite of a well-trained model. Considering the wide variety of medical records, a larger data can improve this problem.

In the modality category with the largest number of classes, our model has an improvement rate of 15.2% in accuracy compared to the baseline. The starting value of the loss function of the modality in FIGURE 10 is higher than that of yes-no, indicating that the fitting of the modality data is more difficult. The loss value of the modality on the validation set is around 0.4, similar to yes-no, showing that our proposed model has a strong fitting ability. The accuracy of modality exceeds yes-no, indicating that the distribution of modality test data is more consistent with the training data than that of yes-no data.

For relatively balanced category (yes-no), the effect of data enhancement is not significant. The data enhancement we used is an elementary transformation based on the existing image. This method only makes the model not inclined to output a certain class, but the improvement of data diversity is limited. For example, ''ugi-upper gi'', which accounts for only 0.4% in the training set, gets 30% accuracy in the test set after data enhancement. The proportion has been increased, but it is still the lowest.

Similar to the results given by [9], open-ended abnormality is difficult to achieve high accuracy. The answers predicted in the generative mode are phrases such as ''glioblastoma multiforme''. These words are related to the type of images and the word frequency in the training set. High frequency predictions account for 22.6%. The evaluation metrics can only compare which method is better, but they do not fully reflect whether the predicted answer is close to the ground truth. As can be seen in FIGURE 11, the predicted answers and the ground truths of questions have similar expressions and the same meaning, but the answers only get 0.135 and 0.111 BLEU scores respectively. It indicates that when the predicted answer is close to ground truth, it may not get a high score, which is unreasonable. This is the case in 8.8% of the predictions. In addition, 20.2% ground truths include words that cannot be found in the training set. It means that there

are differences in the data distribution between the training set and the test set, and it is impossible to generate the same answer as a ground truth no matter how hard we trained.

There is no medical experts in the process of evaluating the automatic VQA model. And the evaluation metrics are limited. Although strong in fitting, the CGMVQA is still far from being a human doctor. The model has higher accuracy on the elementary questions, so it could be used for assisting teaching beginning medical students or giving the answers to the elementary questions from the patients. Expanding the amount of data can make the model perform better.

VI. CONCLUSION

Computer-aided diagnosis can alleviate the current state of medical resource imbalance in some areas, and medical images are increasingly being employed in medically assisted diagnosis. In this work, we propose the CGMVQA for answering corresponding questions based on medical images. Unlike other work, our model is not restricted to a single disease and can be used for several types of medical images and organs.

Specifically, we use the ImageCLEF 2019 VQA-Med as our data set. We split the data into 5 categories to simplify the complex problem and propose a comprehensive model, including classification and answer generation capabilities. Due to the limited amount of data, we adopt data augmentation on images and tokenization on texts. We use pre-trained ResNet152 model to extract image features and a global average pooling strategy to unify the dimensions of these features. We add token, segment and position embeddings layers together to deal with texts. Special tokens, image and text features are concatenated as the input of our model. We employ the pre-layer-normalization multi-head self-attention transformer to avoid the warm-up optimizer. And we reduce the parameters and share the embedding weight to ensure that the model can be implemented on a single GPU.

We only use images and questions to do the training and classify them by the output of the ''[CLS]'' position. The classification mode is suitable for other categories except abnormality. This mode is influenced by data imbalance. We use images, questions and masked answers to generate answers to the abnormality category. The generative mode predicts the sequence by looping. The result of the generative mode is limited by the amount of data.

We adopt strict accuracy, word matching and semantic similarity as the evaluation metrics. Our model gets results of 0.640 accuracy score, 0.659 BLEU score, 0.678 WBSS score and achieves state-of-the-art results on the ImageCLEF 2019 VQA-Med data set.

There is still a lot of work to be done to apply it in a clinical context. Some diseases may become invisible when we resize images of a large size to 224×224. Next, we will look for more effective data augmentation methods to see if we can achieve a better performance. Existing evaluation metrics do not fully reflect the results, we will try other more realistic metrics or cooperate with experts to involve in the evaluation process and get some suggestions for the improvement of our model.

The code of our proposed model is available on: https://github.com/youngzhou97qz/CGMVQA.

REFERENCES

[1] J. Ker, L. Wang, J. Rao, and T. Lim, ''Deep learning applications in medical image analysis,'' *IEEE Access*, vol. 6, pp. 9375–9389, 2018.

[2] K. Doi, ''Computer-aided diagnosis in medical imaging: Historical review, current status and future potential,'' *Computerized Med. Imag. Graph.*, vol. 31, nos. 4–5, pp. 198–211, Jun. 2007.

[3] R. Wang, X. Liang, X. Zhu, and Y. Xie, ''A feasibility of respiration prediction based on deep bi-LSTM for real-time tumor tracking,'' *IEEE Access*, vol. 6, pp. 51262–51268, 2018.

[4] F. Ren, X. Kang, and C. Quan, ''Examining accumulated emotional traits in suicide Blogs with an emotion topic model,'' *IEEE J. Biomed. Health Informat.*, vol. 20, no. 5, pp. 1384–1396, Sep. 2016.

[5] C. Li, G. Zhu, X. Wu, and Y. Wang, ''False-positive reduction on lung nodules detection in chest radiographs by ensemble of convolutional neural networks,'' *IEEE Access*, vol. 6, pp. 16060–16067, 2018.

[6] D. Bardou, K. Zhang, and S. M. Ahmad, ''Classification of breast cancer based on histology images using convolutional neural networks,'' *IEEE Access*, vol. 6, pp. 24680–24693, 2018.

[7] C. Eickhoff, I. Schwall, A. G. S. de Herrera, and H. Müller, ''Overview of ImageCLEFcaption 2017-image caption prediction and concept detection for biomedical images,'' in *Proc. CLEF, Working Notes*, 2017, pp. 1–10.

[8] S. A. Hasan, Y. Ling, O. Farri, J. Liu, H. Müller, and M. Lungren, ''Overview of ImageCLEF 2018 medical domain visual question answering task,'' in *Proc. CLEF, Working Notes*, 2018, pp. 1–8.

[9] J. J. Lau, S. Gayen, A. Ben Abacha, and D. Demner-Fushman, ''A dataset of clinically generated visual questions and answers about radiology images,'' *Sci. Data*, vol. 5, no. 1, Dec. 2018, Art. no. 180251.

[10] A. Ben Abacha, S. A. Hasan, V. V. Datla, J. Liu, D. Demner-Fushman, and H. Müller, ''VQA-Med: Overview of the medical visual question answering task at imageclef 2019,'' in *Proc. CLEF Work. Notes*, Lugano, Switzerland: CEUR, Sep. 2019, pp. 1–11.

[11] S. J. Pan and Q. Yang, ''A survey on transfer learning,'' *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2009.

[12] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[13] J. L. Elman, ''Finding structure in time,'' *Cognit. Sci.*, vol. 14, no. 2, pp. 179–211, 1990.

[14] S. Hochreiter and J. Schmidhuber, ''Long short-term memory,'' *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[15] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus, ''Simple baseline for visual question answering,'' 2015, *arXiv:1512.02167*. [Online]. Available: http://arxiv.org/abs/1512.02167

[16] T.-Y. Lin, A. RoyChowdhury, and S. Maji, ''Bilinear CNN models for fine-grained visual recognition,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1449–1457.

[17] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, ''Learning phrase representations using RNN encoder-decoder for statistical machine translation,'' 2014, *arXiv:1406.1078*. [Online]. Available: http://arxiv.org/abs/1406.1078

[18] F. Ren and J. Deng, ''Background knowledge based multi-stream neural network for text classification,'' *Appl. Sci.*, vol. 8, no. 12, p. 2472, 2018.

[19] P. Wang, Q. Wu, C. Shen, A. van den Hengel, and A. Dick, ''Explicit knowledge-based reasoning for visual question answering,'' 2015, *arXiv:1511.02570*. [Online]. Available: http://arxiv.org/abs/1511.02570

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, ''Attention is all you need,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ''BERT: Pre-training of deep bidirectional transformers for language understanding,'' 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805

[22] Y. Wu *et al.*, ''Google's neural machine translation system: Bridging the gap between human and machine translation,'' 2016, *arXiv:1609.08144*. [Online]. Available: http://arxiv.org/abs/1609.08144

[23] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, ''BLEU: A method for automatic evaluation of machine translation,'' in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2001, pp. 311–318.

[24] G. So ancıo lu, H. Öztürk, and A. Özgür, "BIOSSES: A semantic sentence similarity estimation system for the biomedical domain," *Bioinformatics*, vol. 33, no. 14, pp. i49–i58, Jul. 2017.
[25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[26] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.
[27] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.
[28] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," 2018, *arXiv:1802.05365*. [Online]. Available: http://arxiv.org/abs/1802.05365
[29] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," 2019, *arXiv:1909.11942*. [Online]. Available: http://arxiv.org/abs/1909.11942
[30] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. (2018). *Improving Language Understanding by Generative Pre-Training*. [Online]. Available: https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/languageUnderstand.paper.pdf
[31] Anonymous, "On layer normalization in the transformer architecture," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–17. [Online]. Available: https://openreview.net/forum?id=B1x8anVFPr
[32] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," 2019, *arXiv:1908.02265*. [Online]. Available: http://arxiv.org/abs/1908.02265
[33] H. Tan and M. Bansal, "LXMERT: Learning cross-modality encoder representations from transformers," 2019, *arXiv:1908.07490*. [Online]. Available: http://arxiv.org/abs/1908.07490
[34] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "VL-BERT: Pre-training of generic visual-linguistic representations," 2019, *arXiv:1908.08530*. [Online]. Available: http://arxiv.org/abs/1908.08530
[35] L. Harold Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "VisualBERT: A simple and performant baseline for vision and language," 2019, *arXiv:1908.03557*. [Online]. Available: http://arxiv.org/abs/1908.03557
[36] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2425–2433.
[37] A. Di Martino *et al.*, "The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism," *Mol. Psychiatry*, vol. 19, no. 6, pp. 659–667, 2014.
[38] A. Suhr, S. Zhou, A. Zhang, I. Zhang, H. Bai, and Y. Artzi, "A corpus for reasoning about natural language grounded in photographs," 2018, *arXiv:1811.00491*. [Online]. Available: http://arxiv.org/abs/1811.00491
[39] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi, "From recognition to cognition: Visual commonsense reasoning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6720–6731.
[40] Y. Zhou, X. Kang, and F. Ren, "Tua1 at ImageCLEF 2019 VQA-med: A classification and generation model based on transfer learning," in *Proc. CEUR Workshop*, vol. 2380. Aachen, Germany: RWTH Aachen, 2019.
[41] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017.
[42] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6077–6086.
[43] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," 2017, *arXiv:1712.04621*. [Online]. Available: http://arxiv.org/abs/1712.04621
[44] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
[45] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, *arXiv:1312.4400*. [Online]. Available: http://arxiv.org/abs/1312.4400
[46] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*. [Online]. Available: http://arxiv.org/abs/1606.08415
[47] S. Wiseman and A. M. Rush, "Sequence-to-sequence learning as beam-search optimization," 2016, *arXiv:1606.02960*. [Online]. Available: http://arxiv.org/abs/1606.02960
[48] M. F. Medress, F. S. Cooper, J. W. Forgie, C. C. Green, D. H. Klatt, M. H. O'Malley, E. P. Neuburg, A. Newell, D. Reddy, and B. Ritea, "Speech understanding systems: Report of a steering committee," *Artif. Intell.*, vol. 9, no. 3, pp. 307–316, 1977.
[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[50] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.
[51] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *Proc. 32nd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 1994, pp. 133–138.
[52] X. Yan, L. Li, C. Xie, J. Xiao, and L. Gu, "Zhejiang university at ImageCLEF 2019 visual question answering in the medical domain," in *Proc. CEUR Workshop*, vol. 2380. Aachen, Germany: RWTH Aachen, 2019.
[53] E. Loper and S. Bird, "NLTK: The natural language toolkit," 2002, *arXiv:cs/0205028*. [Online]. Available: https://arxiv.org/abs/cs/0205028

FUJI REN (Senior Member, IEEE) received the Ph.D. degree from the Faculty of Engineering, Hokkaido University, Japan, in 1991. From 1991 to1994, he worked as a Chief Researcher at CSK. In 1994, he joined the Faculty of Information Sciences, Hiroshima City University, as an Associate Professor. Since 2001, he has been a Professor with the Faculty of Engineering, Tokushima University. His current research interests include natural language processing, artificial intelligence, affective computing, and emotional robot. He is a Fellow of The Japan Federation of Engineering Societies, IEICE, and CAAI. He is the Academician of The Engineering Academy of Japan and EU Academy of Sciences, a Vice President of CAAI, and the President of International Advanced Information Institute, Japan. He is an Editor-in-Chief of *International Journal of Advanced Intelligence*.

YANGYANG ZHOU received the B.E. degree from Zhejiang University, China, in 2015. He is currently pursuing the master's and Ph.D. degree with Tokushima University, Japan. His research interests include computer vision and natural language processing.