Visual Question Generation and Answering for Radiology Images

Bhavin Jain [1] Janani Sankarasubramanian [1] Sahil Sawant [1]

Abstract

We are interested in generating questions and answers from medical images(radiographs) like generating natural language questions based on the contents of radiology images. For the generation of the questions and answers. We first generated new training data from the existing examples, based on contextual word embeddings and image augmentation techniques. It then uses the variational auto-encoders model to encode images into a latent space and decode natural language questions. For the answer generation task, we fine-tuned BERT with Masked Lan-guage Modelling as the pretext task on ROCO dataset. This gave us state-of-the-art results on both VQA-RAD and Image Clef 2019 datasets.

1. Introduction

The motivation behind this paper is ImageCLEF VQA-MED challenge. In 2020, they presented the participants with two tasks, Visual question Generation and Visual Question Answering. In this paper, we explore these two tasks on VQA-RAD and ImageCLEF 2020 datasets.

Visual Question Answering(VQA) on medical radiology images intends to construct models that can answer natu-ral language questions asked on medical images by provid-ing the medical professionals useful information as well as assisting patients in interpreting their medical images. In the areas of Natural Language Processing, Computer Vi-sion and Language, self-supervised pretraining of BERT architectures has proven to be very effective.

Generating questions based on the contents of image in the medical field is a less explored area particularly for the radiology images. The scarcity of large amount of labeled data in the medical area makes supervised learning ap-proaches ineffective for training purposes. Thus we want to perform data augmentation technique on the images to increase the size of a training dataset by producing changed versions of the images in the dataset and the questions to overcome the data restriction problem of medical Visual Question Generation(VQG). We propose a model based on the variational auto-encoders(VAE) architecture and is de-signed to take a radiology image as input and output natural question and answer pairs.

The paper is organized as follows: Section 2 Related work. Section 3 is the proposed VQG method, data augmenta-tion techniques and the detailed visual question generation and answering methodology. Section 4 and 5 presents the datasets and experimental results. Section 6 and 7 describes the discussions and conclusion.

1.1. Research contributions

Most datasets focus on a single medical condition, for example, lung cancer. We acheive start of the art results in VQA-Med2019 dataset that covers 36 image modalities including CT, MR, etc, it covers 10 distinct orgrans, and various abnormalities. We are able to beat the first place solution at the leaderboard and with the help of VQA-Med, we can now look at a variety of conditions and we hope to help physicians and patients in diagnosis by establishing a question and answer system.

A deep learning model generally works well when it has a huge amount of data. So, we attempt to create a sys-tem that can generate questions based on radiology images based on the VAE architecture and data augmentation to automatically generate new training sets.

2. Related Work

The VQG task was introduced for the first time in Im-ageCLEF 2020 VQA-MED challenge. VQG in the open-domain benefited from the available large annotated datasets (Agrawal et al., 2015). There is a variety of work studying generative models for generating visual questions in the open domain (Mora & de la Puente, 2016). Recent VQG approaches have used variational autoencoders archi-tecture and it is proven to be successful (Jain et al., 2017). Due to the problem of data scarcity in medical VQG, We artificially generate new training data. In this paper, we present a new VQG system capable of generating questions

[1]Worcester Polytechnic Institute. Correspondence to: Janani Sankarasubramanian <jsankarasubraman@wpi.edu>.

CS/DS 541 Deep Learning Final Project, Worcester Polytechnic Institute.

about radiology images. The system is primarily based on the VAE architecture and data augmentation.

Since the first medical VQA challenge was being organized in 2018 (Hasan et al., 2018), there were increasing number of researches that made medical VQA an inspiring field. BERT-like architectures have proven to be effective in Nat-ural Language Processing. Our solution is to use a multi-modal transformer architecture for medical VQA.

3. Proposed Method

The goal of this study is to generate natural language questions from radiology images and to use the gener-ated dataset of sufficient size as a training set for VQA tasks. A shift in image data distribution might result in sub-optimal performance when using pre-trained weights from general domain. Hence model needs to trained with medi-cal domain knowledge base rather than a general language database. These factors motivate the need for learning se-mantic representations of medical images and texts from scratch. Owing to the attention operation, we use Trans-former encoder for learning effective representations.

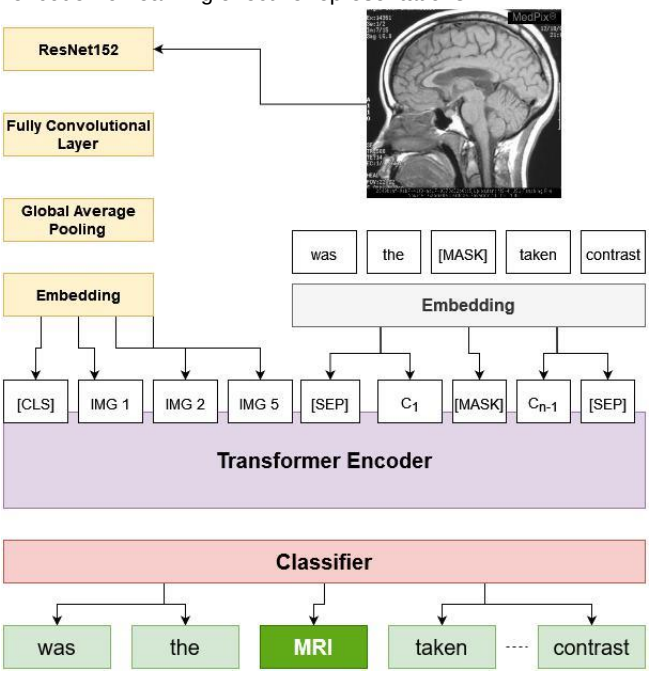

Fig 1: This is the pre-training part for Visual Question An-swering. Here you can observe that the medical terminol-ogy is masked before sending to Transformer Encoder, and as a result we get MRI as an output to replace the MASK token

3.1. Data Augmentation

3.1.1. Images

We generated new training instances based on image aug-mentation techniques. To do so, we applied flipping, ro-tation, shifting, and blurring techniques to all VQA-RAD training images. For pretraining and finetuning, we resized all images to 224 × 224. We use image crop, rotation and color jitter for image data augmentation.

3.1.2. Questions

We first select nouns and verbs as candidate words, using the part-of-speech tags NN, NNS, NNPS, NNP, VBD, VBP, VBN, VBG, VBZ, VB. Each candidate word is then re-placed by contextually similar words using Wiki-PubMed-PMC embedding which was trained using four million En-glish Wikipedia, PubMed, and PMC articles.

3.2. Visual Question Generation

The proposed VQG system is based on the variational au-toencoders architecture (Kingma & Welling, 2013). It first encodes the image before generating the question. VAEs consist of two neural network modules, encoder, and de-coder, for learning the probability distributions of data p(x). The encoder creates a latent variable z from raw data x and transforms it into latent space z space. The decoder plays the role of recovering x using z extracted from the latent space. Given an image v, a CNN is used for obtaining a feature map and encoding the dense vectors h_v into a latent (hidden) representation z-space. It then reconstructs the in-puts from the z-space using a simple Multi Layer Percep-tron (MLP) which is a neural network with fully connected layers. Finally, it uses a decoder LSTM to generate the question $q\hat{}$ from the z-space.

3.3. Visual Question Answering

3.3.1. Pre-Training

The image features are extracted from ResNet152 and passed through an embedding layer. The caption is tok-enized and the keywords are masked with [Mask] tokens. The text embeddings are obtained by combining input, po-sition and segment embeddings. The final embedding is passed through a transformer encoder. The encoder outputs are then passed to a classifier which predicts the masked words. We use BERT WordPiece tokenizer (Devlin et al., 2019) for text tokenization. The sequence of 5 image fea-tures and the caption token embeddings together are pro-vided as input to the BERT- like model. Unlike BERT-BASE (Devlin et al., 2019) our model has 4 BERT Layers and a total of 12 attention heads.To ensure that the model learns to predict medical words from the context, we mask

only medical keywords (provided with the dataset) from the keywords and leave the common words untouched.

3.3.2. Fine-Tuning

We load the model with weights from pre-training and fine-tune it further on the train split of the respective medical VQA dataset. Instead of using [CLS] (Classification) to-ken representation from the last layer of the Transformer, we average the representation of each token obtained from the last layer and further pass it through dense layers for classification.

4. Experiment

4.1. Datasets

In this study, we used the VQA-RAD dataset (Lau, 2019) of clinical visual questions and images. It contains 315 im-ages and 3,515 corresponding questions of 11 types. Each image is associated with more than one questions. Radiol-ogy Objects in Context (ROCO) (Pelka et al., 2018) dataset contains over 81,000 radiology images with several medi-cal imaging modalities. For pretraining, we use all the im-ages, their corresponding captions and use the keywords for masking. VQA-Med 2019 (Ben Abacha et al., 2019) is a challenge dataset introduced as part of the ImageCLEF-VQA Med 2019 challenge. It contains radiology images and has four main categories of questions: Modality, Plane, Organ system and Abnormality. All the samples having Yes/No as the ground truth are considered as Yes/No cat-egory. The dataset includes a training set of 3200 medical images with 12,792 Question-Answer (QA) pairs, a valida-tion set of 500 medical images with 2000 QA pairs and a test set of 500 medical images with 500 QA pairs.

4.2. Visual Question Generation

We implemented the model using PyTorch. We used ImageNet-pretrained ResNet-152 (He et al., 2015) pro-vided by PyTorch as the image encoder and did not fine-tune its weights. LSTM decoder is used for generating questions. All images are resized to 224*224. Adam opti-miser with a learning rate of 0.0001 and a batch size of 8 is used. All models are trained for 40 epochs and the best validation results are used as final results

4.3. Visual Question Answering

A model was pre-trained on ROCO and fine tuned on all samples in the train split of the respective VQA dataset. Then we fine tuned the same model using the weights from pre-training to separately train the model for different question categories like MODALITY, PLANE, ORGAN, ABN(abnormality), etc. We limited the question types to

5. For pre-training, we optimize the loss using Adam opti-mizer with learning rate $2e^{-5}$ and reduce the learning rate by a factor of 0.1 if the validation loss does not improve for 5 consecutive epochs. For fine tuning, we use the Adam op-timizer, learning rate of $1e^{-4}$ and reduce the learning rate by a factor of 0.1 if validation loss does not improve for 10 consecutive epochs

5. Results

VQG in medical domain is a very challenging task. We applied some data augmentation techniques to im-prove the metrics. But we did not achieve the expected results. Here we use some language modelling met-rics such as BLEU score, METEOR, Rouge-L, CIDEr.

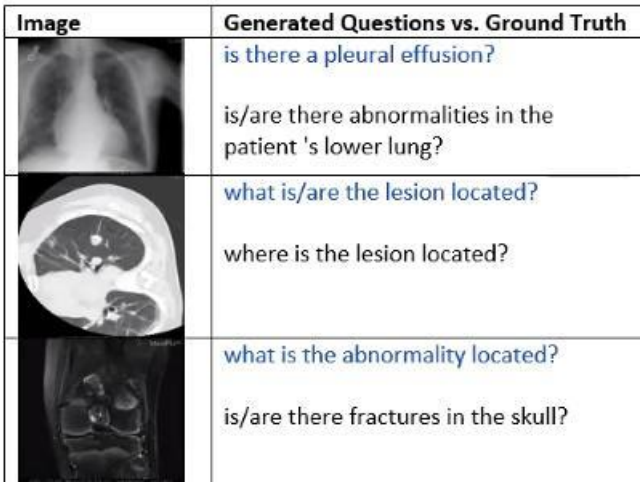

Fig 2: Here we can see the generated questions are in blue and the expected questions or the ground truths are in black. We use Accuracy and BLEU score to evaluate the VQA performance. Our model achieves state of the art results even surpassing the first posi-tion in the leaderboard of ImageCLEF 2019 challenge.

5.1 Visual Question Generation

Bleu_1	Bleu_2	Bleu_3	Bleu_4	METEOR	ROUGE_L	CIDER
0.2325	0.1115	0.0445	0.0191	0.0968	0.2998	0.1265

5.2 Visual Question Answering

Dataset	Test Accuracy	Test BLEU score
VQA-MED 2019	60%	0.6124

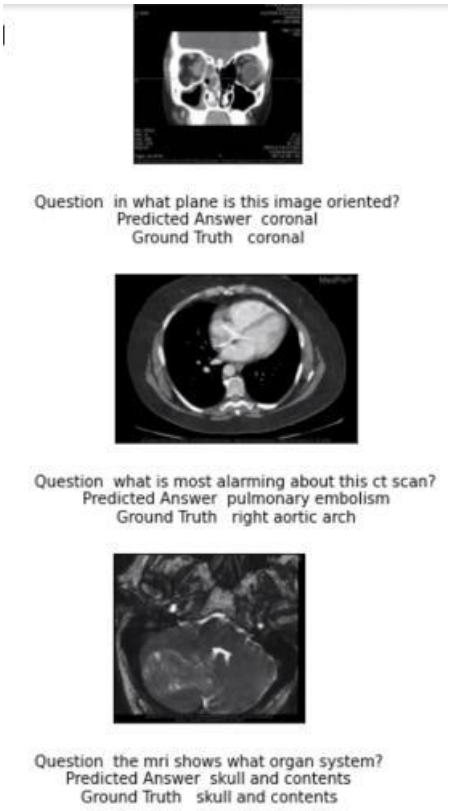

Fig 3: Here we can see the questions and the ground truths along with predicted answers.

6. Discussion

Right now our models are trained only for a limited set of data and fall short to answer questions in detail. Most of the questions being answered are in yes/no format or in couple of words. However, we feel with a more detailed dataset in future from ImageCLEF we can make our model perform even better in both domains.

7. Conclusions and Future Work

The results of the BLEU score and manual evaluations showed that VQG under-performs by generating relevant but few repetitive questions. We learned that the weights and right dataset plays a key role in generating questions. However, with VQA we were able to outperform many state of the art techniques working with radiology image dataset with a outstanding BLEU score of 0.69. The BERT model plays a crucial role in pre-training by MASKing the critical biological keywords helping to train model efficiently. For future work we will try to re-adjust the weights and gather better labelled data involving human interference(medical experts) which will be helpful in generation of better questions with more contextual meaning.

References

Agrawal, Aishwarya, Lu, Jiasen, Antol, Stanislaw, Mitchell, Margaret, Zitnick, C. Lawrence, Batra, Dhruv, and Parikh, Devi. Vqa: Visual question answering, 2015. URL https://arxiv.org/abs/1505.00468.

Ben Abacha, Asma, Hasan, Sadid A., Datla, Vivek V., Liu, Joey, Demner-Fushman, Dina, and Muller,¨ Henning. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In CLEF 2019 Working Notes, CEUR Workshop Proceedings, Lugano, Switzerland, September 9-12 2019. CEUR-WS.org <http://ceur-ws.org>.

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Hasan, Sadid A., Ling, Yuan, Farri, Oladimeji, Liu, Joey, Muller,¨ Henning, and Lungren, Matthew P. Overview of imageclef 2018 medical domain visual question answer-ing task. In CLEF, 2018.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Jain, Unnat, Zhang, Ziyu, and Schwing, Alexander. Creativity: Generating diverse questions using variational autoencoders. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5415–5424, 2017. doi: 10.1109/CVPR.2017.575.

Kingma, Diederik P and Welling, Max. Auto-encoding variational bayes, 2013. URL https://arxiv. org/abs/1312.6114.

Lau, J. J.and Gayen, S.and Demner D.and Ben Abacha A. Visual question answering in radiology (vqa-rad). Febru-ary 2019.

Mora, Issey Masuda and de la Puente, Santiago Pascual. Towards automatic generation of question answer pairs from images. 2016.

Pelka, Obioma, Koitka, Sven, Ruckert,¨ Johannes, Nensa, Felix, and Friedrich, Christoph M. Radiology objects in context (roco): A multimodal image dataset. In Stoy-anov, Danail, Taylor, Zeike, Balocco, Simone, Sznit-man, Raphael, Martel, Anne, Maier-Hein, Lena, Duong,

Luc, Zahnd, Guillaume, Demirci, Stefanie, Albarqouni, Shadi, Lee, Su-Lin, Moriconi, Stefano, Cheplygina, Veronika, Mateus, Diana, Trucco, Emanuele, Granger, Eric, and Jannin, Pierre (eds.), Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis, pp. 180–189, Cham, 2018. Springer International Pub-lishing. ISBN 978-3-030-01364-6.