SSN MLRG at MEDVQA-GI 2023: Visual Question Generation and Answering using Transformer based Pre-trained Models

Sheerin Sitara Noor Mohamed[1, *], *Kavitha Srinivasan*[2], *Raghuraman Gopalsamy*[3]

[1, 2, 3] *Department of CSE, Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam – 603110, India*

Abstract

The technological development in current era demands the need of Artificial Intelligence (AI) in all fields. The AI in medical field is not an exception for various real time applications as per user demands. The applications are medical report summarization, image captioning, Visual Question Answering (VQA) and Visual Question Generation (VQG). ImageCLEF is one of the forum which constantly conducing the challenges in these applications. In this paper, for the given MEDVQA-GI dataset, three medical VQA and one medical VQG models are proposed. The medical VQA models are developed using VisionTransformer (ViT), SegFormer and VisualBERT techniques through a combination of eighteen QA-pairs based on categories and resulted an accuracy of 95.6%, 95.7% and 62.4% respectively. Also, the proposed medical VQG model is developed using Category based Medical Visual Question Generation (CMVQG) technique only.

Keywords

Medical VQA, Medical VQG, ImageCLEF, Vision Transformer, SegFormer, VisualBERT, Category based Medical Visual Question Generation, QA-pairs

1. Introduction

The Medical Visual Question Answering and Generation is an challenging field in Natural Language Processing (NLP) and Computer Vision because of the complex nature of both image and text. The ImageCLEF [1] , an online evaluation forum analysis the current trends and conducing the research related tasks since 2018. In 2018 [2] and 2019 [3], they concentrate on visual questions related to different organs, planes, modalities and abnormalities. Then in 2020 [4] and 2021 [5], ImageCLEF concentrated on abnormality type questions alone based on the inferences made from previous two years. This year [6], they are conducing a task for colonoscopy based Visual Question Answering (VQA) and Visual Question Generation (VQG).

In VQA and VQG datasets given by ImageCLEF is based on the HyperKvasir dataset and Kvasir Instrument dataset. These datasets are used to develop proposed models using suitable techniques and

[1]CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece
EMAIL: sheerinsitaran@ssn.edu.in (A. 1); kavithas@ssn.edu.in (A. 2); raghuramang@ssn.edu.in (A. 3)
ORCID: 0000-0003-1752-2107 (A. 1); 0000-0003-3439-2383 (A. 2)

evaluated performance metrics, are discussed in the following paragraphs. The VQA approaches are: joint embedding [7], hybrid, compositional and transformer based techniques. Among these techniques, transformer based techniques is chosen because it hold the potential to understand the relationship between sequential elements and performs parallel processing more quickly. The different transformer based techniques used in this paper are, VisionTransformer (ViT) [8], SegFormer [9], VisualBERT [10] and Category based Medical Visual Question Generation (CMVQG for VQG). The reason behind choosing these techniques are, (i). ViT incorporates more global information then other pre-trained model at lower layers, leading to quantitatively different features (ii). VisualBERT is a simple and flexible framework for modelling a broad range of vision and language tasks (iii). SegFormer has its own advantage over the speed, accuracy, number of parameters (iv). The CMVQG generates the questions based on the category instead of answer in the VQA dataset.

The rest of the paper is organized as follows. Section 2 explains about the MEDVQA-GI 2023 task and dataset description. Section 3 briefs the system design of the proposed VQA and VQG models. Section 4 analyses the results using suitable quantitative metric, and Section 5 concludes with the future work.

2. Task and Dataset Description

In this section, two sub tasks of MEDVQA-GI 2023 and given dataset is explained. The two sub tasks includes, Visual Question Answering (VQA) and Visual Question Generation (VQG).

2.1. ImageCLEF MEDVQA-GI 2023 task

ImageCLEF, a part of Conference and Labs of the Evaluation Forum is conducting tasks related to the medical domain since 2018. Its goal is that through the combination of text and image data the output of the analysis gets easier to use by medical experts.

In sub task A (VQA), the answer to the colonoscopy image needs to be generated with respect to the given the question-answer pairs. For example, given the image containing the colon polyp with the question, “What type of polyp is present?”. Then the answer should be textual description of the type of the polyp located in the image.

In sub task B (VQG), the question to the colonoscopy image need to be generated based on the significant informantion present in the image and answer. The significant information includes, location, count, color, size, shape of the polyps, modality, abnormality type, etc.

2.2. ImageCLEF MEDVQA-GI 2023 dataset

The MEDVQA-GI 2023 task consists of two sub tasks namely, VQA and VQG. The image, mask and QA-pairs for both tasks are given in Table 1. Each sub task consists of training set and test set. In these tasks, 18 QA-pairs are associated with each image so the count of QA-pairs is eighteen times the number of images. These questions are tabulated along with the frequent answers for each question and categories in Table 2. Based on these categories, VQA and VQG model is generated and it is discussed in Section 3.

Table 1
Dataset description for MEDVQA-GI 2023 task

Task	Input/Output	Training Set	Test Set
VQA/ VQG	Image	2000	1949
	Mask	500	-
	QA pairs	36000	-

Table 2
Question and its frequent answers in MEDVQA-GI 2023 task

Q.No	Questions	Frequent Answers	Categories
1	Are there any anatomical landmarks in the image?	Z-line, Not relevant, Cecum, No	Classification type QA-pairs
2	How many findings are present?	0, 1, 2	Numeric oriented QA-pairs
3	How many instruments are in the image?	0, 1	Numeric oriented QA-pairs
4	How many polyps are in the image?	0, 1	Numeric oriented QA-pairs
5	Is there a green/black box artefact?	Yes, No	Color oriented QA-pairs
6	Is there text?	Yes, No	Classification type QA-pairs
7	Are there any abnormalities in the image?	Polyps, Oesophagitis, Ulcerative colitis, No	Classification type QA-pairs
8	Is this finding easy to detect?	Yes, No	Location oriented QA-pairs
9	What color is the anatomical landmark?	Not relevant, Pink	Color oriented QA-pairs
10	Are there any instruments in the image?	Not relevant, Tube, No	Classification type QA-pairs
11	Have all polyps been removed?	Not relevant, No	Location oriented QA-pairs
12	What type of polyp is present?	Not irrelevant, Paris ip and Paris iia	Classification type QA-pairs
13	What type of procedure is the image taken from?	Colonoscopy	Classification type QA-pairs
14	Where in the image is the anatomical landmark?	Centre, Not relevant	Location oriented QA-pairs
15	Where in the image is the instrument?	Centre, Not relevant	Location oriented QA-pairs
16	Where exactly in the image is the polyp located?	Imagename_scan.png	Location oriented QA-pairs
17	What color is the abnormality?	Pink, red, white; Pink, red; Not relevant	Color oriented QA-pairs
18	Have all polyps been removed?	Not relevant, No	Location oriented QA-pairs

3. System Design

The system design of the proposed medical VQA and VQG models are shown in Figures 1, 2, 3 and 4. For medical VQA, three models are developed using VisionTransformer (ViT), SegFormer and VisualBERT based on its categories and one medical VQG model created using Category based Medical Visual Question Generation (CMVQG) techniques as given in Table 3.

Table 3
Dataset category description

Techniques	Categories	Justification
ViT	Classification type QA-pairs (6) Numeric oriented QA-pairs (3)	Possible to apply transformer architecture with self-attention for sequence of images without using convolutional layers.
VisualBERT	Location oriented QA-pairs (3)	Capable of representing significant object in an image by a bounding region.
SegFormer	Color oriented QA-pairs (6)	Feasible to do image segmentation by identifying different segment using “binary mask classification”

The three VQA models are developed using the colonoscopy image and QA-pairs in the training set and, is validated by predicting the label for the test set. The Vision Transformer (ViT) VQA model is developed for QA pairs under classification or numeric type and is shown in Figure 1. The ViT divides the input image into patches of 16×16 pixels and linearly projects the flattened patches. The QA-pairs are converted into tokens using patch and position embedding. Based on the patches and tokens, the model is trained autoregressively for predicting the next token under causal (or unidirectional) self-attention using Multi Layer Perceptron (MLP). The model was implemented using the Vision Encoder Decoder class from the Hugging Face Transformers library and tiny Data-efficient image Transformer (DeiT) pre-trained on ImageNet dataset.

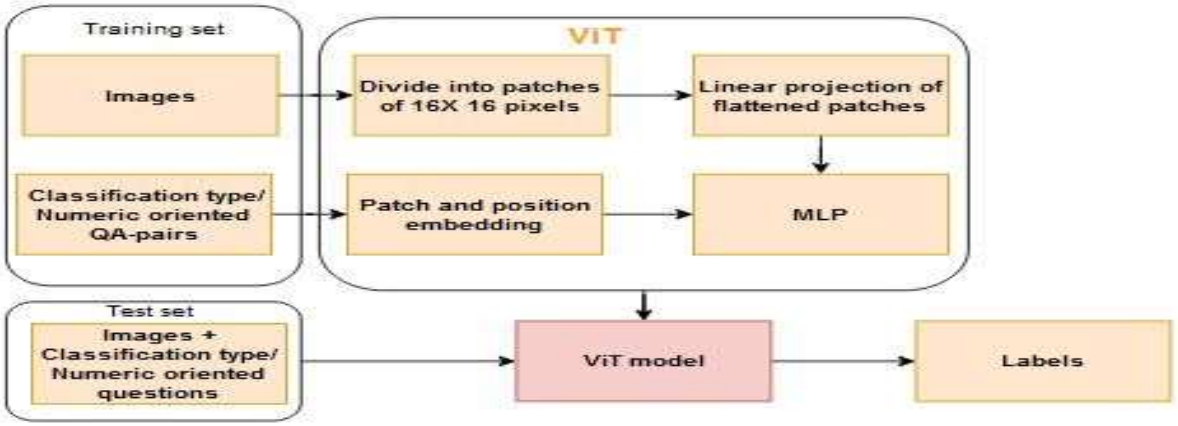

Figure 1: VQA system using ViT technique

The VisuaBERT is an extension of Bidirectional Encoder Representations from Transformers (BERT), model an image with respect to the bounding region in the image. In VisualBERT, the tokens and vocabulary lists are generated using position and segment embedding and it is concatenated with image features to generate model during the training phase. Faster Region based Convolutional Neural Networks (RCNN) is used in order to extract the features from an image and to represent the segmented region with the bounding box. From the segmented region, the appearance features are extracted and is then embedded with the text features to generate the model and it is shown in Figure 2.

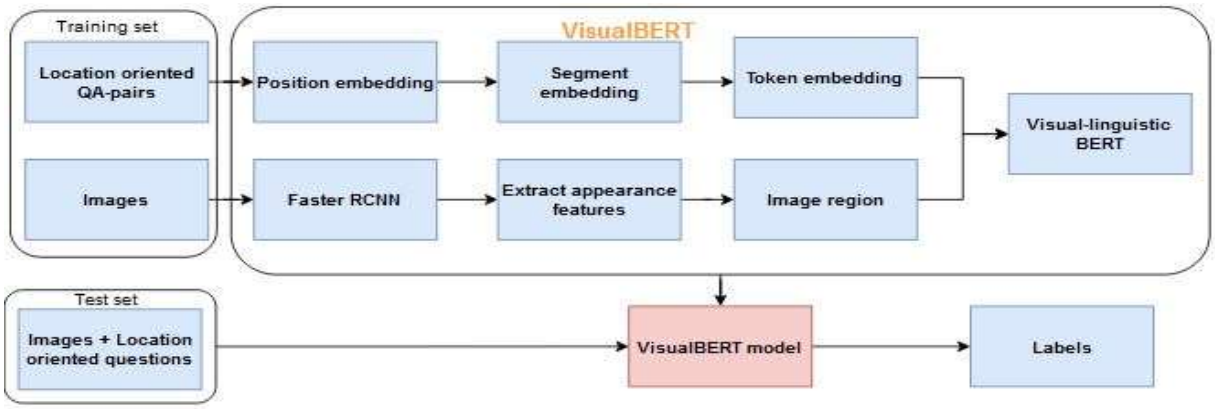

Figure 2: VQA system using VisualBERT technique

In Figure 3, the images were divided into small patches which favor the dense prediction task. These patches are given as input to the attention layer to obtain multi-level features of the original image resolution. It is then passed to Multi-Layer Perceptron to implicitly discover useful alignments between both sets of inputs in terms of color and build up a new joint representation in the training phase. Finally, the generated model is validated by predicting the answers for the color oriented questions for the given colonoscopy image.

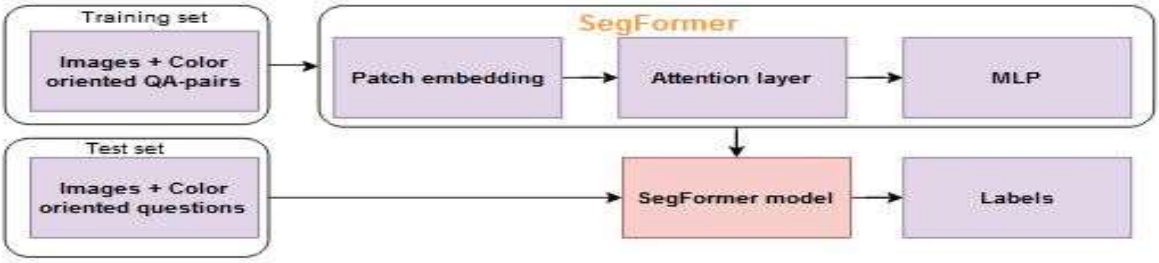

Figure 3: VQA system using SegFormer technique

For medical VQG model generation, Category based Medical Visual Question Generation (CMVQG) approach is used and it is shown in Figure 4. In the proposed CMVQG approach, the Convolutional Neural Network (CNN), Multi-Layer Perceptron (MLP) and Long Short Term Memory (LSTM) are used in the training phase. Because, CNN is capable of extracting the image features and to learn the internal representation of an image. MLP remembers the pattern in the sequential data and is used to extract the text features from the given answers, questions and categories with respect to an image. Finally, LSTM handles long term dependency for extended period of time. Following this, two encoders are used to generate latent encoding from the features of both image as well as text. Later, both the generated latent encoding is concatenated by passing it to the weighted MLP which generates the corresponding latent representation. This concatenated latent representation acts as a backbone that contains the significant information for question generation. The final model is generated by passing this concatenated latent representation to the LSTM and it generates the question as a sequence of words based on the previous words.

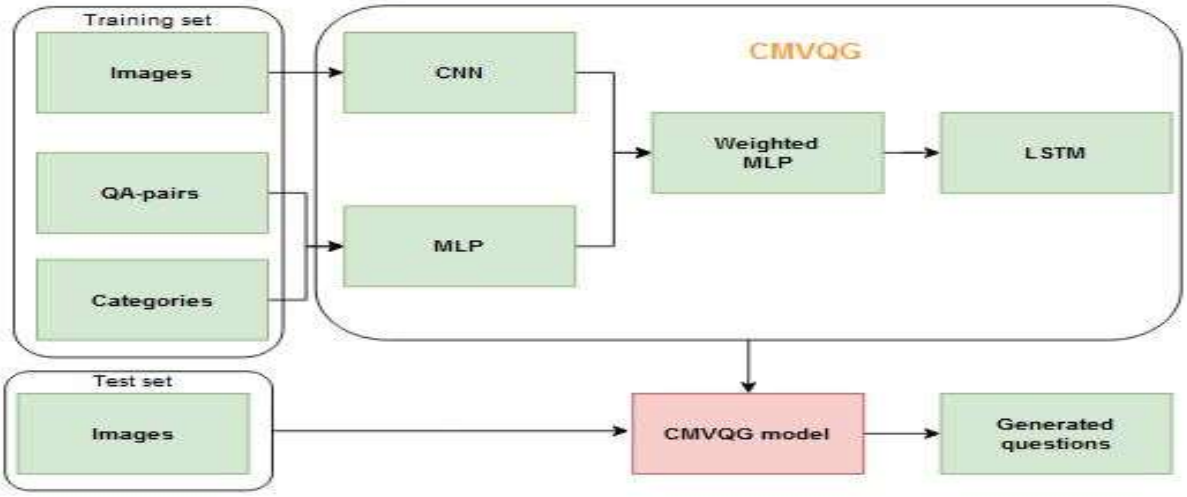

Figure 4: VQG system using CMVQG technique

4. Experiments and Results

The hardware and software required for the implementation of VQA and VQG models includes, (i). Intel i5 processor with NVIDIA GeForce Ti 4800, 4.3GHZ clock speed, 16GB RAM, Graphical Processing Unit and 2TB disk space, (ii). Linux – Ubuntu 20.04 operating system, Python 3.7 package with required libraries like tensorflow, torch, sklearn, nltk, pickle, pandas, etc.

The three VQA models and one VQG model are created and validated using MEDVQA-GI 2023 dataset. The VQA models are initially is trained for 20 epochs, and finally for an additional 20 epochs starting from the checkpoint with the lowest validation loss with an learning rate of 5×10^{-5}. Due to limitations of the computational resources available unable to fine tune the model using self-critical sequence training. Then the performance of VQA models are analyzed for each question and it is given

in Table 4 and 5. In Table 4, it has been inferred that the overall accuracy is 0.471. In addition to this, the classification type QA-pairs attained an highest accuracy of 0. 956. The highest, lowest and overall accuracy for each category is given in Table 5 for better understanding.

Table 4
Brief description about each question in Run 1

Q.No	Question	Category	Accuracy
1	Are there any anatomical landmarks in the image?	Classification type QA-pairs	0.956
2	How many findings are present?	Numeric oriented QA-pairs	0.433
3	How many instruments are in the image?	Numeric oriented QA-pairs	0.604
4	How many polyps are in the image?	Numeric oriented QA-pairs	0.577
5	Is there a green/black box artefact?	Color oriented QA-pairs	0.566
6	Is there text?	Classification type QA-pairs	0.662
7	Is this finding easy to detect?	Location oriented QA-pairs	0.481
8	Are there any abnormalities in the image?	Classification type QA-pairs	0.000
9	What color is the anatomical landmark?	Color oriented QA-pairs	0.957
10	Are there any instruments in the image?	Classification type QA-pairs	0.363
11	Have all polyps been removed?	Location oriented QA-pairs	0.578
12	What type of polyp is present?	Classification type QA-pairs	0.529
13	What type of procedure is the image taken from?	Classification type QA-pairs	0.618
14	Where in the image is the anatomical landmark?	Location oriented QA-pairs	0.624
15	Where in the image is the instrument?	Location oriented QA-pairs	0.527
16	Where exactly in the image is the polyp located?	Location oriented QA-pairs	0.529
17	What color is the abnormality?	Color oriented QA-pairs	0.043
18	Have all polyps been removed?	Location oriented QA-pairs	0.578
	Overall		**0.471**

Table 5
Brief description about range of accuracy under each category

Techniques	Categories	Highest accuracy	Lowest accuracy	Overall accuracy
ViT	Classification type QA-pairs	0.956	0.000	0.521
VisualBERT	Location oriented QA-pairs	0.624	0.481	0.553
ViT	Numeric oriented QA-pairs	0.604	0.433	0.538
SegFormer	Color oriented QA-pairs	0.957	0.043	0.522

From Table 4, it has been inferred that, the VisualBERT and ViT maintains the reasonable accuracy for all types of QA-pairs and hence the accuracy ranges from 40% to 60%. But SegFormer and ViT (for classification Type QA-pairs) are question specific and hence its attains the highest accuracy of 95% as well as the lowest accuracy of 1%.

5. Conclusion

This research experimented the category based approach to solve VQA and VQG tasks of ImageCLEF MEDVQA-GI 2023. For this task, three VQA models such as ViT, SegFormer,

VisualBERT and CMVQG are developed and validated. From the results of the proposed models, it has been inferred that SegFormer and ViT are more problem specific and hence the overall performance is 52.2% and 53.0% respectively which will be improved by choosing the appropriate QA-pairs categories with respect to the medical image. On the other hand, VisualBERT are task generic so it performs reasonably better for mostly all VQA datasets and it ranges from 48.1% to 62.4%. In the future work, the performance can be improved by creating medical related transformer based models. The overall performance can be improved by concentrating on the abnormality type questions.

6. Acknowledgements

Our profound gratitude to Sri Sivasubramaniya Nadar College of Engineering, Department of CSE, for allowing us to utilize the High Performance Computing Laboratory and GPU Server for the execution of this challenge successfully.

7. References

[1] Hasan, S. A., Ling, Y., Farri, O., Liu, J., Müller, H., & Lungren, M. P. (2018, September). Overview of ImageCLEF 2018 Medical Domain Visual Question Answering Task. In *CLEF (Working Notes)*.

[2] Chamberlain, J., Campello, A., Wright, J., Clift, L., Clark, A., & Seco de Herrera, A. G. (2019, July). Overview of ImageCLEFcoral 2019 task. In *CEUR Workshop Proceedings* (Vol. 2380). CEUR Workshop Proceedings.

[3] Ben Abacha, A., Sarrouti, M., Demner-Fushman, D., Hasan, S. A., & Müller, H. (2021). Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain. In *Proceedings of the CLEF 2021 Conference and Labs of the Evaluation Forum-working notes*. 21-24 September 2021.

[4] Ben Abacha, A., Sarrouti, M., Demner-Fushman, D., Hasan, S. A., & Müller, H. (2021). Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain. In *Proceedings of the CLEF 2021 Conference and Labs of the Evaluation Forum-working notes*. 21-24 September 2021.

[5] Bogdan Ionescu, Henning Muller, Ana-Maria Druagulinescu, Wen-wai Yim, Asma Ben Abacha, Neal Snider, Griffin Adams, Meliha Yetisgen, Johannes Ruckert, Alba Garcia Seco de Herrera, Christoph M. Friedrich, Louise Bloch, Raphael Brungel, Ahmad Idrissi-Yaghir, Henning Schafer, Steven A. Hicks, Michael A. Riegler, Vajira Thambawita, Andrea Storås, Pål Halvorsen, Nikolaos Papachrysos, Johanna Schöler, Debesh Jha, Alexandra-Georgiana Andrei, Ahmedkhan Radzhabov, Ioan Coman, Vassili Kovalev, Alexandru Stan, George Ioannidis, Hugo Manguinhas, Liviu-Daniel Stefan, Mihai Gabriel Constantin, Mihai Dogariu, Jerome Deshayes, Adrian Popescu, Overview of ImageCLEF 2023: Multimedia Retrieval in Medical, SocialMedia and Recommender Systems Applications Experimental IR Meets Multilinguality, Multimodality, and Interaction, *Proceedings of the 14th International Conference of the CLEF Association (CLEF 2023)*, Springer Lecture Notes in Computer Science LNCS}, Thessaloniki, Greece.

[6] Steven A. Hicks, Andrea Storås, Pål Halvorsen, Thomas de Lange, Michael A. Riegler, Vajira Thambawita, Overview of ImageCLEFmedical 2023 – Medical Visual Question Answering for Gastrointestinal Tract, CLEF2023 Working Notes, *CEUR Workshop Proceedings*, September 18-21, 2023, Thessaloniki, Greece.

[7] Noor Mohamed, S. S., & Srinivasan, K. A comprehensive interpretation for medical VQA: Datasets, techniques, and challenges. *Journal of Intelligent & Fuzzy Systems*, (Preprint), 1-17.

[8] Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., & Tao, D. (2022). A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, *45*(1), 87-110.

[9] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, *34*, 12077-12090.

[10] Li, L. H., Yatskar, M., Yin, D., Hsieh, C. J., & Chang, K. W. (2019). Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

[11] https://keras.io/examples/vision/image_classification_with_vision_transformer/
[12] https://huggingface.co/blog/mask2former
[13] https://datasets.simula.no/hyper-kvasir/
[14] https://datasets.simula.no/kvasir-instrument/