Project Report on

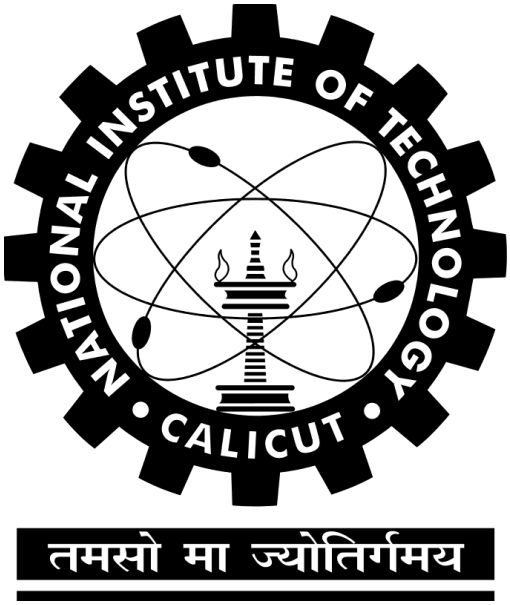
MEDICAL VISUAL QUESTION ANSWERING

Submitted by

Gagan Lal B190480CS
Geethika S B190449CS

Under the Guidance of

Dr. Saidalavi Kalady



Department of Computer Science and Engineering
National Institute of Technology Calicut
Calicut, Kerala, India - 673 601

OCTOBER 11, 2022

Medical Visual Question Answering

Gagan Lal
Dept. of Computer Science and Engineering
National Institute of Technology Calicut
B190480CS

Geethika S
Dept. of Computer Science and Engineering
National Institute of Technology Calicut
B190449CS

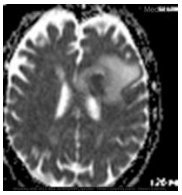
***Abstract*— A VQA system is expected to answer an image related question. This project aims to build a system which predicts the answers to medical questions through the understanding and reasoning of vision (image). When given a medical image and a clinically relevant question, the system is supposed to predict a convincing answer. Having a reliable Med VQA system which can provide a second opinion on medical cases can be really helpful for patients as well as medical professionals. Even though many studies were conducted in this domain, Med VQA still needs more exploration.**

***Key Words* — Visual Question Answering (VQA), Computer vision (CV), Natural Language Processing (NLP), Convolution Neural Network (CNN), Recurrent Neural Network (RNN), Long short-term memory (LSTM)**

1. INTRODUCTION

Lately, A lot of researches are being done in the sub domains of machine learning. With advancements in the fields of natural language processing and computer vision, visual question answering has grasped the attention of the researchers. A VQA system is expected to answer an image related question through the understanding of the provided image and question. For example, given an image of a bird, a person can ask "What is shown in the figure?" or "What is the color of the bird?" The VQA system is expected to provide the correct answer. VQA requires a deep comprehension of both images and textual questions. It is purely supervised learning setting. VQA systems combine NLP, which provides an understanding of the question and the ability to produce an answer, with CV techniques, which provide an understanding of the content of the image.VQA shows great potential in interpreting automated medical imagery and machine supported diagnoses. Since 2016, a general VQA challenge has been issued every year to answer questions of various types. In 2018, ImageCLEF launched VQA-Med challenge which is specific to the medical field and is held annually.

In a Med VQA system, both the image and question given to the system are from the medical domain. Med VQA is technically more challenging than general VQA because of the following factors: (i) creating a large scale Med VQA dataset is challenging because expert annotation is expensive for its high requirement of professional knowledge and QA pairs cannot be synthetically generated directly from images. (ii) Answering questions according to a medical image also demands a specific design of the VQA model. (iii) A question can be very professional, which requires the model to be trained with medical knowledge base rather than a general language. A complete Med VQA system can directly review patients' images and answer any kind of questions. It will ease the shortage of medical resources and provide convenience for patients as well as medical professionals. The most recent model proposed in 2022 [7], has a BLEU values of 0.162, 0.687 and 0.753 for 3 VQA-Med datasets.



Q: What is abnormal in the MRI?

A: A brain abscess, central nervous system

Fig 1. An example of Medical VQA where an image and question are given as input and the system predicts the correct answer.

2. PROBLEM STATEMENT

Given a medical image and a clinically relevant question, give a convincing answer according to the visual clues present in the medical image.

3. LITERATURE REVIEW

A. Methods for VQA

In [1], the authors have focused on studying about the various VQA methods which are divided into four categories based on the nature of their main contribution and datasets available for training and evaluating VQA systems. The methods mentioned in the paper are (i) Joint embedding approaches which allow one to learn representations in a common feature space. Joint embedding model was proposed by H.gao et al. for implementing VQA. It focuses on the global features of an image. Image and textual question are taken as input and features of both are to be extracted through different deep learning and NLP techniques. After getting these features, both the feature vectors are jointly embedded into common feature space and then this combined feature vectors are fed into classifier. The classifier then predicts the answer to the question. (ii) Attention mechanism focuses on question specific region of an image rather than all the global

features of an image. They use local image features and allows model to assign different importance to features from different regions. (iii) Compositional models are useful when questions require multi step reasoning to answer properly. They facilitate transfer learning as same module can be used and trained within different overall architectures and tasks. (iv) Models using external knowledge bases are useful when additional background knowledge or common sense is required to answer properly.

B. Datasets

In [1], the authors did a detailed survey about the datasets available for training and evaluating VQA systems. The general datasets mentioned in the paper are (i) DAQUAR (Dataset for question answering real world images) which was the first VQA dataset to be designed as benchmark. The images in DAQUAR are split to795 training and 654 test images. (ii) COCO-QA dataset includes 123,287 images (72,783 for training and 38,948 for testing). This dataset uses images from the Microsoft Common Objects in Context data. (iii) Freestyle multilingual image question answering dataset uses 123,287 images. (iv) VQA-real is one of the most widely used dataset. VQA-real comprises of 123,287 training and 81,434 test images. (v) Balanced dataset contains 10,295 and 5328 pairs of complementary scenes for the training and test set respectively. (vi) KB-VQA dataset contains questions requiring topic specific knowledge that is present in DBpedia. (vii)FVQA dataset contains only questions which involve external information. It was designed to include additional annotations

In [2], the authors conducted a detailed study about medical VQA datasets. There are 8 public-available medical VQA datasets up to date: (i) VQA-MED-2018 is a dataset that was proposed in the ImageCLEF 2018. In this dataset the QA pairs were generated from captions by a semi-automatic approach. (ii) VQA-RAD is a radiology-specific dataset proposed in 2018. (iii) VQA-MED-2019 was published during the ImageCLEF 2019 challenge. This dataset addressed four most frequent question categories: plane, modality, organ system and abnormality. (iv) RadVisDial is the dataset for visual dialog in radiology. It consists of two datasets: a silver standard and a gold standard dataset. (v) PathVQA explores VQA for pathology. The questions are designed according to pathologist certification of the American Board of Pathology. (vi) VQA-MED-2020 was published in ImageCLEF 2020 challenge. The images are selected with limitation that the diagnosis was made according to image content. The questions are specifically addressing on abnormality. (vii) SLAKE have both semantic labels and a structural medical knowledge base. (viii) VQA-MED-2021 is published in ImageCLEF 2021 challenge. The training set is same as those in VQA-MED-2020 but the test is new.

C. Medical Visual Question Answering

[3] proposes using meta learning approach to overcome the scalability issues faced when VQA is trained by using large training set of example questions, images and answers. Meta learning approach implies that the model learns to learn i.e. it learns to use set of examples provided at test time to answer the given question. The model is initially trained on a small set of questions/answers and is provided with support set of additional examples at test time. The model proposed in the paper by the authors is basically a deep neural network which takes advantage of meta learning scenario. The approach mentioned is to provide the model with supervised data at test time. The conclusions drawn from this experiment was that even though the baseline is most effective with frequent answers, the proposed model fares better in long tail of rare answers. The proposed model had better sample efficiency and unique capability to learn to produce novel answer. This model improved practicality and scalability of the system.

[4] mainly focuses on type aware medical visual question answering. Since medical images may restrict to specific part of human body, identifying type of the image helps to successfully exploit the characteristics of image. The authors proposed an image feature extraction module which extracts type point. The textual features are joined with type point embeddings. This improved the ability of semantic alignment between modalities. Further, it enhanced the applicability of fusion method for Med VQA. The model achieves state-of-art with VQA-RAD dataset.

[5] proposes an optimal deep neural network based model for answering visual medical question. The medical questions were classified based on a BERT model. The authors used EfficientNet as a deep learning model to extract visual features. These features were then combined using an attention model. Adaptive generic algorithm model was used to determine the optimal learning parameters. This model performed better than the runs of ImageCLEF 2019 participants.

[6] discusses hybrid deep learning model for answering visual medical questions. The medical questions were classified based on the BERT model. The features of medical image were extracted by a hybrid deep learning model of VGG and ResNet. The text features were extracted using a Bi-LSTM mode. The authors combined these features extracted on a classifier based on soft max layer and got the most accurate answer. The dataset used for this model was ImageCLEF2019.

In [7], the authors proposed a bi branch model for medical VQA. The first branch transformer structure is as the main framework of parallel structure mode. A parallel network was adopted to extract the image features. An improved CNN model was used to extract spatial features of the medical

images. Then RNN model was used to extract the sequence features of the medical images. The text features are also extracted. The image features are then embedded into the front part of the question (text) features. These two features are integrated into a feature matrix and then input into the stacked four layer transformer structure. As a result, the model has the capability to learn about the dependency between image features and question features and capture the internal structure of the input vector. In the second branch, the answers of the training set are used as labels of the corresponding images. This model achieves state-of-art-result with 3 datasets – ImageCLEF2018, ImageCLEF2019, VQA-RAD. The main metric score exceeds the best results so far by 0.2%, 1.4% and 1.

TABLE I
TABULAR COMPARISON OF LITERATURE SURVEY

Paper	Method Used	Pros	Cons
[3]	A deep neural network which takes advantage of meta learning scenario.	Model had better sample efficiency and unique capability to learn to produce novel answer.	Handling the memory of dynamic weights was not done properly
[4]	TI,TQ (Type Image, Type Question) modules exploits characteristics of input data	Achieves state-of-art with VQA-RAD , Very high accuracy	Restricted to specific class, not a complete solution for VQA
[5]	BERT model on Question and feature extraction using a Bi-LSTM mode EfficientNet as a deep learning model to extract visual features	Performed better than runs of ImageCLEF 2019 participants, Very High accuracy rate.	Information provided in the questions and the corresponding images are not always sufficient to predict the right answer, and answering the questions often requires external knowledge resources.
[6]	Hybrid deep learning model of VGG and ResNet on ImageBERT model on Question and feature extraction using a Bi-LSTM mode	On using various optimization algorithms on ImageCLEF2019 dataset -- Adam and SGD performed better	Better question classification system needed. Abnormality question answering is poor.
[7]	CNN+RNN on Image. Three layer word embedding based on a biomedical corpus on Text	Achieves state-of-art result of 3 VQA-Med (ImageCLEF2018, ImageCLEF2019, VQA-RAD) datasets. Main metric score exceeds the best results so far by 0.2%, 1.4%, 1.1%.	In the first branch of the BPI-MVQA model, image features and text features are simply connected and then input into the transformer structure model, which indicates that we still lack adequate multi-modal feature fusion

4. WORK DONE

We did a literature survey on the methods and datasets for Med VQA. This initial study focused on learning about the various datasets available and understanding them.

5. WORK PLAN

We aim to develop a medical visual question answering system which focuses on answering questions related to abnormalities of a human body part. We intend to build a model which allows better utilization of abnormality features. For achieving this, we have to choose one dataset from the datasets available which is appropriate for the problem. We aim to do the feature extraction of both image and question using CNN and RNN. CNN can be used to capture the relevant features in the image model which could give a better accuracy. We will be focusing on developing a model that would outperform the already existing models.

6. CONCLUSION

This project aims to develop a system which predicts the answers to medical questions through the understanding of the medical image provided. Med VQA can enhance confidence in diagnosing diseases and helping patients better understand their medical conditions. It is important that the system provides accurate information in minimal time. Even though many researches and studies were conducted in this domain, there is a

space for improvement. The Med VQA model we intend to build focuses on answering questions related to abnormalities of a human body part. The existing works have low accuracy. We will be focusing on developing a model that has more accuracy.

7. REFERENCES

[1] Gupta, A.K., 2017. Survey of visual question answering: Datasets and techniques. *arXiv preprint arXiv:1705.03865*.

[2] Lin, Z., Zhang, D., Tac, Q., Shi, D., Haffari, G., Wu, Q., He, M. and Ge, Z., 2021. Medical visual question answering: A survey. *arXiv preprint arXiv:2111.10056*. Inchur, Vilas and L S, Praveen and Shankpal, Preetham. (2020).

[3] Teney, D. and van den Hengel, A., 2018. Visual question answering as a meta learning task. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 219-235)

[4] Zhang, A., Tao, W., Li, Z., Wang, H. and Zhang, W., 2022, May. Type-Aware Medical Visual Question Answering. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4838-4842). IEEE.

[5] Gasmi, K., Ltaifa, I.B., Lejeune, G., Alshammari, H., Ammar, L.B. and Mahmood, M.A., 2022. Optimal deep neural network-based model for answering visual medical question. *Cybernetics and Systems*, *53*(5), pp.403-424.

[6] Gasmi, K., 2022. Hybrid deep learning model for answering visual medical questions. *The Journal of Supercomputing*, pp.1-18.

[7] Liu, S., Zhang, X., Zhou, X. and Yang, J., 2022. BPI-MVQA: a bi-branch model for medical visual question answering. *BMC Medical Imaging*, *22*(1), pp.1-1

[8] Hasan, S.A., Ling, Y., Farri, O., Liu, J., Müller, H., Lungren, M.P., 2018. Overview of ImageCLEF 2018 medical domain visual question answering task., in: CLEF (Working Notes).

[9] Ben Abacha, A., Datla, V.V., Hasan, S.A., Demner-Fushman, D., Müller, H., 2020. Overview of the VQA-Med task at ImageCLEF 2020: Visual question answering and generation in the medical domain, in: CLEF 2020 Working Notes, CEUR-WS.org, Thessaloniki, Greece.

[10] 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: CLEF 2021 Working Notes, CEUR-WS.org, Bucharest, Romania.