Project Report on

MEDICAL VISUAL QUESTION ANSWERING

Submitted by

Gagan Lal	**B190480CS**
Geethika S	**B190449CS**

Under the Guidance of

Mr. 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:

Medical Visual Question Answering (MED VQA) aims to predict 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 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.

***Index Terms*—VQA, Med VQA, NLP, CV**

1. INTRODUCTION

Visual Questing Answering (VQA) is an interdisciplinary research problem that has drawn extensive attention recently. It incorporates computer vision (CV) and natural language processing (NLP). The VQA system is expected to answer an image-related question through the understanding of the provided image. It is supposed to ease the shortage of medical resources and provide convenience for patients as well as medical professionals. A complete Med VQA system can directly review patients' images and answer any kind of questions.

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 computer vision 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.

A Med VQA system is made up of three major tasks: (i) Medical question classification, (ii) Image and text feature extraction, (iii) Retrieval and extraction of answers in textual form. In Med VQA the images are taken with medical equipment for a specified part of the human body. Hence this is very different from the general VQA task. Medical images can be classified by parts of human body. The image of different parts of the human body corresponds to different questions. 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.

2. PROBLEM STATEMENT

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

3. LITERATURE REVIEW

3.1. VQA: a survey of methods and datasets

Paper Citation: A.K. Gupta, Survey of visual question answering: datasets and techniques, arXiv preprint arXiv:1705.03865 (2017).

Basically in VQA, the computer is presented with an image and a textual question about this image. The computer is supposed to determine the correct answer. VQA needs multimodal knowledge beyond a single sub-domain since it requires information not present in the image. Hence it constitutes a truly AI complete task. This survey mainly focuses on various VQA methods through four categories based on the nature of their main contribution and datasets available for training and evaluating VQA systems.

3.1.1. Methods for VQA

3.1.1.1. Joint Embedding Approaches

They were motivated by the advances of deep neural networks in both computer vision and NLP which allow one to learn representations in a common feature space. Joint embedding model was proposed by H.gao *et al*. for implementing VQA. 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. This model focuses on the global features of an image.

3.1.1.2. Attention Mechanisms

Most of the models use global features to represent the visual input which may feed irrelevant information to the prediction stage. In order to overcome this, attention mechanisms focus 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.

3.1.1.3. Compositional Models

These 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.

3.1.1.4. Models using external knowledge bases

These models are useful when additional background knowledge or common sense is required to answer properly. Wang *et al*. proposed a VQA framework named “Ahab” that uses DBpedia. Visual concepts are first extracted from the given image with CNNs, and they are then associated with nodes from DBpedia. The final answer is obtained by summarizing the results of mapping of images/questions to queries.

3.1.2. Datasets

3.1.2.1. Datasets of natural images

DAtaset for QUestion Answering on Real-world images – DAQUAR is the first VQA dataset designed as benchmark. It was built with images from the NYU-Depth v2 dataset (Silberman *et al.*), which contains 1449 RGBD images of indoor scenes, together with annotated semantic segmentations. The images of DAQUAR are split to 795 training and 654 test images.

The COCO-QA dataset uses images from the Microsoft Common Objects in Context data (COCO) dataset (Lin *et al.*). COCO-QA includes123, 287 images (72,783 for training and 38,948 for testing). This dataset (Ren *et al.*) represents a substantial effort to increase the scale of training data for VQA.

Freestyle Multilingual Image Question Answering) dataset (Gao et al.) – FM-IQA uses 123,287 images. Unlike COCO-QA the questions/answers are provided here by humans through the Amazon Mechanical Turk crowd-sourcing platform which lead to much greater diversity of questions.

VQA-real, (Antol *et al.*), one of the most widely used dataset comprises of two parts, one using natural images named VQA-real, and a second one with cartoon images named VQA-abstract. VQA-real comprises 123,287 training and 81,434 test images.

3.1.2.2. Dataset of clipart images

VQA abstract scenes consists of clipart scenes with question/answer pairs as a separate and complimentary set to real images which enables research focused on high level reasoning. It contains a total of 50,000 scenes and 3 questions per scene.

Balanced dataset - The authors in Zhang et al. (2016) balance the existing abstract binary VQA dataset with additional complementary scenes. The resulting balanced dataset contains 10,295 and 5328 pairs of complementary scenes for the training and test set respectively.

3.1.2.3. Knowledge base-enhanced datasets

The KB-VQA dataset (Wang et al. 2015) was constructed to evaluate the performance of the Ahab VQA system (Wang et al. 2015). It contains questions requiring topic-specific knowledge that is present in DBpedia.

The FVQA dataset (Wang et al., 2016) contains only questions which involve external (non-visual) information. It was designed to include additional annotations to ease the supervised training of methods using knowledge bases.

3.2. VQA as Meta Learning Task

Paper Citation: Damien Teney, Anton van den Hengel : "Visual Question Answering as a meta learning task", 2018

Since VQA is typically trained by using large training set of example questions, images and answers, there will be scalability issues because it attempts to represent all world knowledge within finite set of parameters. In order to overcome this issue, Meta learning approach is proposed. This basically 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 which is most probably big. The model is supposed to exploit the additional data without the need for retaining the model. This greatly improves practicality and scalability of the system.

The model proposed is basically a deep neural network which takes advantage of Meta learning scenario. The approach is to provide the model with supervised data at test time. The model extends the state of art VQA system of Tenet *et al.* During training, the support set is sub sampled to yield a set of 1,000 elements.

The conclusions drawn from the experiment was that even though the baseline is most effective with frequent answers, the proposed model fares better (mostly positive values) in the long tail of rare answers. The proposed model had better sample efficiency and a unique capability to learn to produce novel answer.

3.3. BPI – MVQA: a bi branch model for medical visual question answering

Paper Citation: Shengyan Liu , Xuejie Zhang, Xiaobing Zhou and Jian Yang : BPI_MVQA: a bi-branch model for medical visual question answering

Bi-branched model – Parallel networks and image retrieval – BPI-MVQA

First branch transformer structure is as the main framework of parallel structure mode. We adopt a parallel network to extract the image features. Firstly, we use an improved CNN model to extract the spatial features of the medical images. Secondly, we use an RNN model to extract the sequence features of the medical images. Text feature is also extracted. We embed the above mentioned image features into the front part of the question(text) features, integrate the two parts of features into a feature matrix and then input it into the stacked four-layer transformer structure. As a result, the model can learn the dependency between image features and question features and capture the internal structure of the input feature vector.

Second branch we can use the answers of the training set as the labels of the corresponding images, ignoring the influence of the question features on the classification results.

Achieves state-of-art result of 3VQA – Med (ImageCLEF2018, ImageCLEF2019,VQA-RAD) datasets. Main metric score excess by 0.2%, 1.4%, 1.1%

3.4. Hybrid Deep Learning model for answering visual medical questions

Paper Citation: Karim Gasmi , Accepted: 19 March 2022 / Published online: 11 April 2022 : Hybrid deep learning model for answering visual medical questions.

The classification of medical questions based on a BERT model extraction of medical image features by a hybrid deep learning model of VGG and ResNet Text feature extraction using a Bi-LSTM mode. By combining features extracted on classifier based on softmax layer we get the most accurate answer.

On using various optimization algorithms on ImageCLEF2019 dataset – Adam and SGD performed better.

3.5. Type Aware Medical Visual Question Answering

Paper Citation: Anda Zhang, Wei Tao, Ziyan Li, Haofen Wang, Wenqiang Zhang ,Academy for Engineering and Technology, Fudan University, Shanghai, China, College of Design and Innovation, Tongji University, Shanghai, China : Type-aware Medical Visual question answering

Medical images may restrict to specific part of human body which result in type effect. By identifying type of image we can successfully exploit the characteristics of image. Our image feature extraction module now extracts one more thing called type point. We join textual features with type point embeddings and do VQA. This improves the ability of semantic alignment between modalities and further enhances the applicability of the fusion method for Med VQA

Achieves state-of-art with VQA-RAD

3.6. Optimal Deep Neural Network Baed Model for Answering Visual Medical Question

Paper Citation: Karim Gasmi, Ibtihel Ben Ltaifa, Gaël Lejeune, Hamoud Alshammari, Lassaad Ben Ammar & Mahmood A. Mahmood (2022) Optimal Deep Neural Network-Based Model for Answering Visual Medical Question, Cybernetics and Systems, 53:5, 403-424, DOI: 10.1080/01969722.2021.2018543

The classification of medical questions based on a BERT model. EfficientNet as a deep learning model is used to extract visual features. Text Feature extracted using Bi-LSTM. Combined features using an attention model we used an adaptive genetic algorithm to determine the optimal deep learning parameters

Performed better than runs of ImageCLEF 2019 participants

TABULAR COMPARISON OF LITERATURE SURVEY

Paper	Method used	Pros	Cons
[13]	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 by 0.2%,1.4%,1.1%	In the first branch of th e BPI-MVQA model, image fe atures and text features are simply connected an d then input into the tra nsformer structure model , which indicates that we still lack adequate multi-modal feature fusion
[12]	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 poorly.
[14]	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.
[10]	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

4. WORK DONE

We did a literature survey based on the existing techniques and datasets for Med VQA and analysed the pros and cons of each methods. In addition we understood the major challenges faced in the domain of VQA. We invested time in understanding and researching about the different datasets available.

5. WORK PLAN

We intend to explore various datasets available for medical VQA and choose one dataset appropriate for the problem. We will also study about the various methods which already exist for question feature extraction and image feature extraction such as CNN, RNN and LSTM. We also have to do a detailed survey of techniques used in medical VQA and find problems of the existing systems and solutions that can be used to resolve them.

5. CONCLUSION

The purpose of this literature review was to study about the researches in the domain of Med VQA. Many studies and discussions were conducted on the different datasets and various techniques used to enhance the performance of the Med VQA system. More research and testing are required to gain a better understanding of these techniques. A detailed survey should be conducted to find problems in the existing systems and solutions to resolve them.

6. REFERENCES

[1] Agrawal, A., Kembhavi, A., Batra, D., Parikh, D.: C-vqa: A compositional split of the visual question answering (vqa) v1. 0 dataset. arXiv preprint arXiv:1704.08243(2017)

[2] Damien Teney, Anton van den Hengel, Australian Institute for Machine Learning – University of Adelaide : Visual Question Answering As a Meta Learning Task (2018)

[3] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, Anton van den Hengel Australian Centre for Visual Technologies, School of Computer Science, The University of Adelaide, Adelaide, SA 5005, Australia : Visual question answering : A survey of methods and datasets

[4] Stanislow Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv batra, C. Lawrence Zitnik, Devi Parikh; Prceedings of the IEEE International conference on computer vision,2015, pp. 2425-2433 : VQA-Visual Question Answering

[5] Chen, K., Wang, J., Chen, L.-C., Gao, H., Xu, W., Nevatia, R., 2015a. ABC-CNN: An attention based convolutional neural network for visual question answering. CoRRabs/1511.05960.

[6] Lu, J., Yang, J., Batra, D., Parikh, D., 2016. Hierarchical question-image co-attention for visual question answering. In: Advances In Neural Information Processing Systems, pp. 289–297

[7] A.K. Gupta, Survey of visual question answering: datasets and techniques, arXiv preprint arXiv:1705.03865 (2017).

[8] S.A. Hasan, Y. Ling, O. Farri, J. Liu, M. Lungren, H. Müller, Overview of the ImageCLEF 2018 medical domain visual question answering task, in: CLEF2018

[9] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, A. van den Hengel, Visual question answering: a survey of methods and datasets, Comput. Vision Image Understanding 163 (2017) 21–40.

[10] Anda Zhang, Wei Tao, Ziyan Li, Haofen Wang, Wenqiang Zhang ,Academy for Engineering and Technology, Fudan University, Shanghai, China, College of Design and Innovation, Tongji University, Shanghai, China : Type-aware Medical Visual question answering

[11] Jitender Singh, Dwarikanath Mahapatra,Deepti R Bathula, Indian Institute of Technology Ropar : Medaical VQA: Mixup helps it keep simple.

[12] Karim Gasmi , Accepted: 19 March 2022 / Published online: 11 April 2022 : Hybrid deep learning model for answering visual medical questions.

[13] Shengyan Liu , Xuejie Zhang, Xiaobing Zhou and Jian Yang : BPI_MVQA: a bi-branch model for medical visual question answering

[14] Karim Gasmi, Ibtihel Ben Ltaifa, Gaël Lejeune, Hamoud Alshammari, Lassaad Ben Ammar & Mahmood A. Mahmood (2022) Optimal Deep Neural Network-Based Model for Answering Visual Medical Question, Cybernetics and Systems, 53:5, 403-424, DOI: 10.1080/01969722.2021.2018543

[15] Kafle, K., and C. Kanan. 2016. Answer-type prediction for visual question answering. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4976–4984. doi: 10.1109/CVPR.2016.538