# Adversarial Learning to Improve Question Image Embedding in Medical Visual Question Answering

Kaveesha Silva
*Department of Computer Science and Engineering*
*University of Moratuwa*
Sri Lanka
silvamkc.17@cse.mrt.ac.lk

Thanuja Maheepala
*Department of Computer Science and Engineering*
*University of Moratuwa*
Sri Lanka
thanuja.17@cse.mrt.ac.lk

Kasun Tharaka
*Department of Computer Science and Engineering*
*University of Moratuwa*
Sri Lanka
kasunt.17@cse.mrt.ac.lk

Thanuja D. Ambegoda
*Department of Computer Science and Engineering*
*University of Moratuwa*
Sri Lanka
thanujaa@uom.lk

2022 Moratuwa Engineering Research Conference (MERCon) | 978-1-6654-8786-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/MERCon55799.2022.9906168

***Abstract*—Visual Question Answering (VQA) is a computer vision task in which a system produces an accurate answer to a given image and a question that is relevant to the image. Medical VQA can be considered as a subfield of general VQA, which focuses on images and questions in the medical domain. The VQA model's most crucial task is to learn the question-image joint representation to reflect the information related to the correct answer. Medical VQA remains a difficult task due to the ineffectiveness of question-image embeddings, despite recent research on general VQA models finding significant progress. To address this problem, we propose a new method for training VQA models that utilizes adversarial learning to improve the question-image embedding and illustrate how this embedding can be used as the ideal embedding for answer inference. For adversarial learning, we use two embedding generators (question–image embedding and a question-answer embedding generator) and a discriminator to differentiate the two embeddings. The question-answer embedding is used as the ideal embedding and the question-image embedding is improved in reference to that. The experiment results indicate that pre-training the question-image embedding generation module using adversarial learning improves overall performance, implying the effectiveness of the proposed method.**

***Keywords*—medical visual question answering, adversarial learning.**

## I. Introduction

Interest in automated medical image interpretations is increasing as the field of artificial intelligence advances. As a result, medical VQA is getting a lot of attention in the medical community since it has the ability to improve clinical decision-making and help patients learn more about their health problems through medical imaging.

The medical VQA system has various advantages for both medical professionals and patients. These types of systems can be used to make clinical decisions or even clarify medical professionals' decisions. Another issue is that medical specialists who can evaluate a medical image and make a diagnosis are limited and often overworked. As a solution using a medical VQA system, only the images that are considered critical can be directed to specialists, thereby saving time for specialists. And patients with access to medical images can learn more about their medical issues via the medical VQA system. And patients can only make an appointment if there is an issue identified by the VQA system. Furthermore, some scans provide a vast number of images, but professionals only look at the ones that are relevant to the patient's symptoms. Other generated photos can be evaluated using an automated medical VQA system in these instances to identify if there is an anomaly early on.

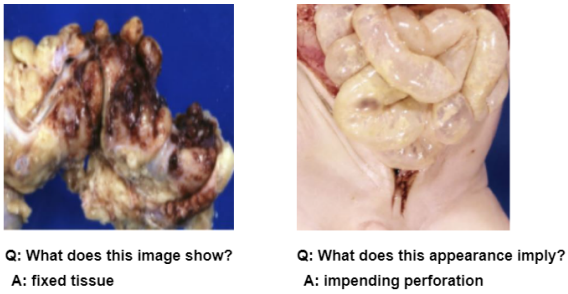


Fig. 1. Examples for situations where the question provides no information about the object or region that is relevant to the answer. (from pathVQA data set)

VQA systems are becoming more accurate and efficient as deep learning and image processing areas develop. However, medical VQA systems are not yet reliable enough to use in actual clinical settings. As previously mentioned, there are lots of advantages of the medical VQA system that can have a significant impact on the medical field. Despite the potential benefits, building a medical VQA system has unique

978-1-6654-8786-3/22/$31.00 ©2022 IEEE

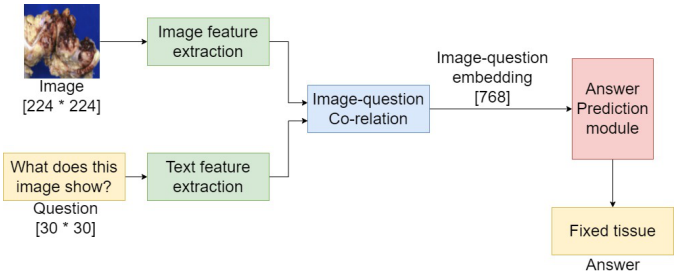


Fig. 2. Main components of a typical VQA Model. Image and the questions are the inputs for the two feature extraction components. The question-image co-relation module generates fused output (vector of size 768 in this scenario) using the image and question features.

challenges. We try to address some of these issues in this work.

In medical VQA, learning the joint representation of question-image pairs is one of the most important tasks. Most existing general VQA models use multi-modal correlation methods to learn joint representation [1], [2]. But, for a variety of reasons, general VQA models do not perform well on medical VQA tasks, hence researchers are attempting to develop VQA models that are specific to medical VQA. In this paper, we proposed a new model that uses adversarial learning methods to learn the question-image joint representation in addition to the multi-modal correlation methods.

## II. Related Work

As the baseline of our experiment, we used the method proposed in [3] for PathVQA [4] data set. In that method, the authors followed a model proposed in LXMERT [5]. Text features are extracted with BERT [6] which uses the splitting and tokenizing approach in [7]. A Faster-RCNN network pre-trained on a medical dataset (A dataset that includes images of blood cells) was used to collect image features. A cross-model encoder was used to learn the relationship between image and text embeddings. That encoder contains self-attention sub-layers, cross-attention sub-layers, and some feed-forward layers. In our architecture, we used the concept of GAN (Generative Adversarial Networks) [8] to improve correlation learning. A GAN consists of two main sub-networks. The discriminator network and the generator network. The generator is responsible for generating better data representations and the discriminator is responsible for discriminating against data created by the generator. In this conceptual framework, both the generator and the discriminator improve similarly to a min-max two-player game. In this study, the authors used adversarial networks to generate images from random noise that cannot be distinguished from training example images.

In our proposed method, similar adversarial learning concepts are used to train a question-image embedding generator that can emulate an ideal embedding that reflects answer information. Instead of starting from random noise, image and question features can be used to generate an improved question-image embedding to infer the answer.

In our literature survey, we reviewed VQA models and training approaches presented in [9]- [13]. To our best knowledge, the possibility of using the adversarial network in medical VQA has not yet been explored in other medical VQA studies. Most of the VQA models are mainly focusing on multi-modal correlation methods to capture the answer-related information in the question-image embedding.

ALMA [14] is a method proposed for visual question answering with adversarial learning. A pre-trained VGG19 network [15] has been used for image feature extraction and a Glove [16] for text feature extraction. Also, the ALMA architecture consists of a Siamese network to learn the correlation between the image and the question. The concept of GAN is used to improve correlation learning and obtain a better question-image embedding. The authors have proposed a framework for integrating adversarial networks into multi-modal correlation learning and have performed experiments on three general VQA data sets. The experimental results indicate improvement, suggesting the effectiveness of adversarial learning. However, all experiments are done in the general VQA domain, which has larger data sets compared to medical VQA data sets. In contrast, our experiments are carried out on the PathVQA medical data set, which represents a wide range of medical concepts with a limited number of data examples.

## III. Problem Definition

As shown in Fig. 2, the four main components of a basic VQA model are the image feature extraction module, question feature extraction module, question-image embedding module, and answer prediction module. Image input and the question (text input) are fed into the image feature extraction module and the question feature extraction module respectively. The output from each module will then be fed into the question-image embedding module, and the question-image embedding module's embedded output will be used to predict the answer.

As indicated in Fig. 2, the image embedding that is used to predict the answer is considerably smaller in size than the original image and question representations. As a result, the joint embedding of the image and question only carries a limited amount of data. To correctly predict the answer, information relevant to the answer should be available in the question-image embedding. Because of this, the most challenging and crucial task in VQA is learning the joint representation of the question-image pair for answer inference.

Usually, general VQA models do not perform well for medical VQA mainly due to, the size of the medical VQA data sets is very low compared to general VQA data sets, and the knowledge base required to answer medical VQA questions is very diverse, yet only a few examples are available for particular medical concepts.

Furthermore, most existing methods focus on analyzing the multi-modal correlation between the question and the image to infer answer-related information. Multi-modal correlation methods attempt to focus more on the image regions present in the question. The majority of the general VQA questions involve an object or a region in the image. As a result, multi-modal correlation approaches will capture the answer-related information for those types of questions effectively.

However, when the question does not include any information about the object or region that is relevant to the answer, these methods may not be able to capture the information adequately. This scenario is illustrated by two questions in Fig. 1. In medical VQA, these types of questions are quite common. As a result, answer-related information could be missing from the question-image joint embedding, resulting in poor medical VQA performance.

In the proposed method, we are focusing on improving the question image embedding to reflect the answer-related information. In this approach, we use adversarial methods to improve question-image joint representation in addition to multi-modal correlation methods.

## IV. Methodology

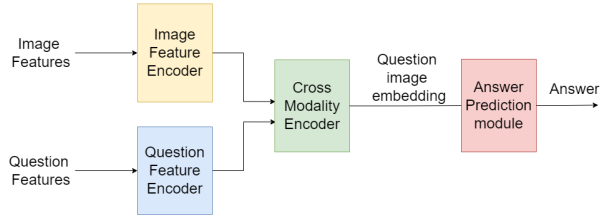


Fig. 3. The architecture of the baseline models with image, question, and cross-modality encoder. Answer prediction module receives the question image embedding as the input. Question image embedding should contain the answer-related information to successfully predict the answer.

A VQA training example consists of an image, a question related to the image, and the answer to the question. The objective of the VQA system is to predict the answer using the image and the question. In the process, the VQA system generates a question-image embedding using the question-answer pairs. Each question and answer pair relates to a specific semantic scene of the image. The answer is linked to that specific semantic scene, and to predict the answer this specific semantic scene is expected to be included in the question image joint embedding. As an example, in Fig. 1 the question "What does this appearance imply?" and the answer "impending perforation" corresponds to the semantic scene "Impending perforation appearances are present in this image". If this semantic scene is encoded into the question-image embedding answer can be retrieved from it.

In the question and answer pair of the previous example, the same semantic scene can be found. The question and answer

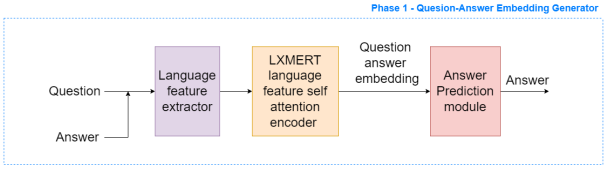


Fig. 4. In the initial phase VQA model is trained using only the language features extracted by the concatenated answer and question string. Question answer embedding generated by the LXMERT is used as the input of the answer prediction module.

in Fig. 1 create a semantic scene as "This appearance implies impending perforation". The intended semantics that needs to be encoded in the question-image embedding is similar to this semantic scene in the question-answer pair. That is question-answer embedding can act as the ideal embedding to infer the answer. As a result, emulating the question-answer embedding with the question-image embedding will allow the question-image embedding to reflect the answer-related information.

To make a question-image embedding similar to a question-answer embedding, our method uses adversarial learning to pre-train the question-image embedding generator.

In our method, we are using the Medical VQA model used in Pathological Visual Question Answering [3] as the baseline. As shown in Fig. 3 the VQA model we are using consists of four main components; an image feature encoder for processing the image features, a language encoder to process the question features, a cross-modal encoder for question image embedding generation, and the answer prediction module to predict the answer from the predefined answer space. In this research, our main focus is on improving the question-image embedding generator. Our proposed training methodology can be divided into 3 phases.

1) Training the model using question and answer
2) Training the question-image embedding generator using adversarial learning
3) Use the pre-trained question-image embedding generator in phase two, in the VQA model and train retraining the entire model.

And our model consists with following components:

- Question-answer embedding generator - language self attention layers in the LXMERT [5]
- question-image embedding generator - cross-modal relationship encoder
- Answer prediction module - a fully connected neural network

### A. Phase 1 - Training question-answer embedding generator

In this phase, we trained the baseline VQA model using the question-answer pairs. To generate the question-answer embedding, we used the language self-attention layers in the LXMERT model as the question-answer embedding generator. For each question-answer pair, we created the input by concatenating the answer and question using the [answer]_[question] pattern. Let us denote one training example as $E$ which has a question $E_q$, an image $E_i$ and an answer $E_a$. For inputs $E_q$ and $E_a$, $X_{E_{qa}}$ is the output generated from the question-answer embedding generator. After training, question-answer embedding $X_{E_{qa}}$ was able to achieve a high level of performance since it has the answer-related information directly encoded. Hence it is possible to use the question-answer embeddings as the ideal embeddings in phase 2. Results of the model performance when using the question-answer embedding are shown in table I. After the training process, the question-answer embedding generator module was saved as a separate module to be used in the adversarial learning phase.

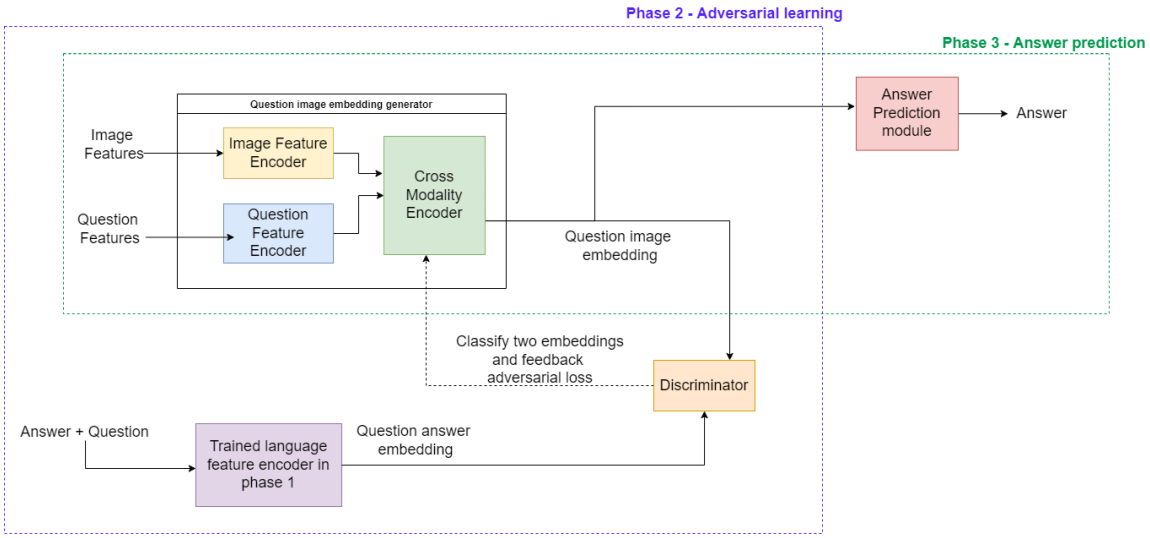


Fig. 5. The high-level architecture diagram of the solution is shown here. The blue dotted box contains the components used in phase 2. To generate the question-answer embedding, the question-answer embedding generator trained in phase 1 is used. The inputs to the discriminator are the question image embedding and the question-answer embedding. The discriminator attempts to distinguish between two embeddings and feeds the adversarial loss back to the question image embedding generator. The green dotted box contains the components utilized in phase 3. The answer is predicted using the output from the improved question image embedding.

TABLE I
ACCURACY OF DIFFERENT TYPES OF OPEN-ENDED QUESTIONS IN PHASE 1

| **Method** | **Question types** | | | | | | |
|---|---|---|---|---|---|---|---|
| | ***Yes/No*** | ***What*** | ***Where*** | ***How*** | ***How much/many*** | ***Why*** | ***Overall Accuracy*** |
| Baseline model | 0.861 | 0.22 | 0.73 | 0.12 | 0.45 | 0.50 | 0.576 |
| **Baseline model with question answer embedding** | **1.00** | **0.67** | **0.94** | **0.34** | **0.82** | **0.68** | **0.838** |

### *B. Phase 2 - Adversarial Learning*

In this phase, the goal is to train the question-image embedding generator, to provide an output as close as possible to that of the question-answer embedding generator $X_{E_{qa}}$, for a given image ($E_i$), question ($E_q$), and answer ($E_a$). Generative adversarial learning [8] is used to train the question image embedding generator. In this process, we introduced a neural network to act as the discriminator. For the question-image embedding generator, we use the LXMERT cross-modal relationship encoder with an additional neural layer. As seen in Fig. 5 the discriminator receives inputs from the trained question-answer embedding generator and the question image embedding generator. The output from the trained question-answer embedding generator is considered as the ground truth. The discriminator tries to classify the two embeddings. The question-image embedding generator will try to confuse the discriminator by emulating the question-answer embedding (the ground truth). In the learning process, both the discriminator and the question image embedding generator will be trained. Two embeddings will be mostly similar when adversarial learning has reached a point of convergence.

Let $D(x)$ denotes the discriminator and $G$ denotes the question-image embedding generator. For a given embedding $x$, $D(x)$ outputs a scalar value indicating the probability of $x$ being a question-answer embedding ($X_{E_{qa}}$). $G$ outputs question-image embedding ($X_{E_{qi}}$) with the inputs $E_q$ and $E_i$. Both $D$ and $G$ are trained simultaneously for each $E$, and training loss will be calculated separately for $G$ and $D$ as $G\_loss$ and $D\_loss$.

Binary Cross-Entropy Loss ($BCELoss$) function is used to calculate the loss. For a prediction $x$ and target $y$, $BCELoss$ can be described as:

$$BCELoss(x, y) = -(y \log(x) + (1 - y) \log(1 - x)) \quad (1)$$

The goal of the $G$ is to create $X_{E_{qi}}$ such that $D(X_{E_{qi}}) = 1$($D$ identifies $X_{E_{qi}}$ as a $X_{E_{qa}}$). To achieve this $G\_loss$ is calculated as:

$$G\_loss = BCELoss(D(X_{E_{qi}}), 1) \quad (2)$$

The goal of the $D$ is to identify two embeddings correctly, such that $D(X_{E_{qi}}) = 0$ and $D(X_{E_{qa}}) = 1$. Therefore to train the discriminator $D\_loss$ is calculated as:

$$D\_loss = \frac{1}{2} \times (BCELoss(D(X_{E_{qi}}), 0) + BCELoss(D(X_{E_{qa}}), 1)) \quad (3)$$

TABLE II
ACCURACY OF DIFFERENT TYPES OF OPEN ENDED QUESTIONS IN PHASE 3

| Method | Question types | | | | | | |
|---|---|---|---|---|---|---|---|
| | *Yes/No* | *What* | *Where* | *How* | *How much/many* | *Why* | *Overall Accuracy* |
| Baseline method | 0.861 | 0.22 | 0.73 | 0.12 | 0.45 | 0.50 | 0.576 |
| **Baseline method with adversarial learning** | **0.867** | **0.24** | **0.75** | **0.16** | **0.41** | **0.68** | **0.587** |

### C. Phase 3 - Answer prediction with improved embeddings

As the last step, the question image embedding generator of the original model is replaced with the trained question image embedding generator and the entire model is trained including the answer prediction component. The modified model is shown in Fig. 5 inside the green dotted box. PathVQA [4] dataset that we used to train the model consists of 4, 092 unique answers. The final output of the model is a vector with a size of 4, 092, each element indicating the probability of each answer being the correct answer. The answer with the highest probability is selected as the prediction of the model. Since the improved question image embedding generator is able to generate outputs that contain the answer-related information, improved performance is achieved in terms of answer prediction accuracy. The final results are included in the table II.

The model is similar to the baseline model but the question image embedding generator is trained separately using adversarial learning to reflect the answer-related information. Fig. 6 shows how the training loss and validation score change as the training progresses during phase 3.

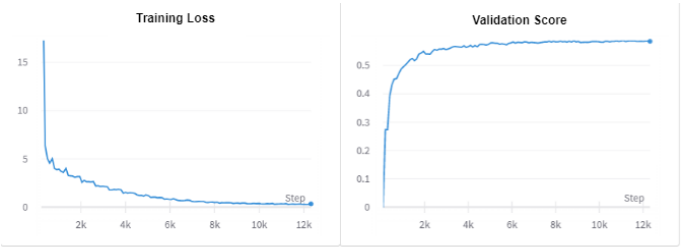


Fig. 6. Training loss and validation score during the training of the entire model in phase 3.

**Implementation details** In the first phase of the research, we trained the baseline VQA model using the question-answer embedding. In training the model we followed the original configurations used in [3]. We used the original data set split of the PathVQA which has a ratio of about 3:1:1 among training, validation, and test set. The batch size was set to 32, the learning rate was 5e-5 and the Adam [17] optimizer was used. The model reached its peak level of performance after training for 20 epochs. The question-answer embedding size was set at 768 to be consistent with the question image embedding. In phase 2, we trained the question-image embedding generator and the discriminator with 5e-5 learning rate, and the Adam [17] optimizer was used. Learning converges after 30 epochs and the batch size used is 32. In this phase, we used a fairly simple discriminator model with five fully connected layers and leaky RELU as the activation function in hidden layers. Hence training time for this phase was mostly dependent on the question-image embedding generator's complexity. In the final phase entire model is trained with the improved question-image embedding generator in phase 2. Same configurations are used as in phase 3 but trained for 40 epochs.

## V. DATA SET

For the model training, we are using the PathVQA data set [4], which includes microscopic images of body tissues, cells, other parts, and relevant question-answer pairs. The data set consists of 4, 998 total images and 32, 799 question-answer pairs. Questions are divided into eight categories. These categories are Yes/No, What, Where, How, How much/many, Why, When and Whose. Both open-ended and closed-ended questions (yes/no answers) are available in the data set.

We used the accuracy [18] metric to evaluate the model. Accuracy measures the percentage of predicted answers that are exactly matching with the ground truth. The data set contains 4, 092 unique answers. The VQA model predicts the probability that each answer is the correct answer for the given question and image pair. The answer with the highest probability is chosen as the prediction.

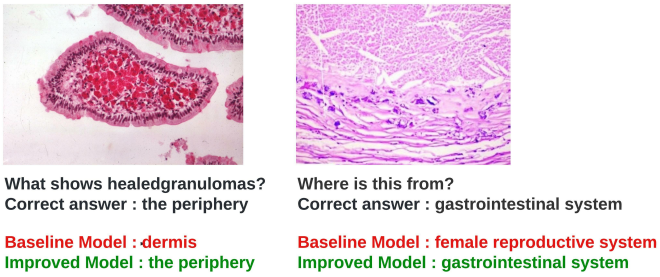


Fig. 7. Two examples of answers predicted by the improved model and the baseline model. The green color represents the correct answers and the red color represents incorrect answers.

## VI. RESULTS

Table II shows the performance comparison between the baseline model and the improved model with the pre-trained question-image embedding generator. We only performed the experiment on PathVQA [4] dataset, but we believe our method should be able to achieve good results on other popular medical VQA data sets such as VQA-Med [19], VQA-Rad

[20], and SLAKE [21]. In the adversarial learning phase, it was difficult to get the losses of two competing models (question-image embedding generator and the discriminator) to a point of convergence. We experimented with different configurations several times to reduce the convergence issue and achieve these results. We can observe overall performance improvement for almost all question types of questions. There is accuracy deterioration for the "How much/many" type of questions. But only 0.4% of questions are included in this category and a few wrong predictions can impact the accuracy results. For all other question types, accuracy improvements can be observed. Fig. 7 shows two examples where the improved model was able to produce correct answers that were previously predicted incorrectly by the baseline model. These findings suggest that it is effective to use the proposed adversarial pre-training method to improve the question-image embedding generations and improve the final performance of the model. However, there is still a significant difference between the performance obtained using question-answer embedding (Table I) and the improved question-image embedding (Table II). This indicates that even with the improved question-image embedding, the answer-related information was not fully captured. We believe that further improvements to the adversarial learning phase of our proposed method could improve performance even more.

## VII. Conclusion

In this paper, we discussed the challenges of creating the Medical VQA model compared to the general VQA model and identified the ineffectiveness of the question-image joint representation to reflect the answer-related information as a limiting factor in improving the performance of medical VQA models. As a solution, we suggest an adversarial learning approach to train question-image embedding generators. We present the experiment results and implementation details of our proposed method. The results of phase 1 (table I) indicate that question-answer embeddings are effective to refer to as the ideal embeddings when training the question image embedding generator. And the final results of phase 3 (table II) show performance improvement over the baseline results. More experiments can be conducted to explore the possibility of using the proposed pre-train approach to improve the performance of other VQA models, including general VQA.

## References


[1] H. Gong, and G. Chen, S. Liu, Y. Yu, G. Li, "Cross-Modal Self-Attention with Multi-Task Pre-Training for Medical Visual Question Answering," 2021, arXiv.org. [Online]. Available: https://arxiv.org/abs/2105.00136.

[2] H. K. Verma and S. Ramachandran, "HARENDRAKV at VQA-Med 2020: Sequential VQA with Attention for Medical Visual Question Answering," 2020, ceur-ws.org. [Online]. Available: http://ceur-ws.org/Vol-2696/paper_62.pdf.

[3] X. He, Z. Cai, W. Wei, Y. Zhang, L. Mou, Eric Xing, et al, "Pathological Visual Question Answering," 2020, arXiv.org. [Online]. Available: https://arxiv.org/abs/2010.12435.

[4] X. He, Y. Zhang, L. Mou, E. Xing, and P. Xie, "PathVQA: 30000+ Questions for Medical Visual Question Answering," 2020, arXiv.org. [Online]. Available: https://arxiv.org/abs/2003.10286.

[5] H. Tan, and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," 2019, arXiv.org. [Online]. Available: https://arxiv.org/abs/1908.07490.

[6] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2019, arXiv.org. [Online]. Available: https://arxiv.org/abs/1810.04805.

[7] Y. Wu, M. Schuster, Z. Chen, Q. Le, M. Norouzi, W. Macherey, et al, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016, arXiv.org. [Online]. Available: https://arxiv.org/abs/1609.08144.

[8] I. Goodfellow, et al, "Generative adversarial nets," in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 2672–2680.

[9] D. Kingma, and J. Ba, "ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION," 2015, arXiv.org. [Online]. Available: https://arxiv.org/abs/1412.6980.

[10] Allaouzi, M. B. Ahmed, and B. Benamrou, "An Encoder-Decoder model for visual question answering in the medical domain," 2019, ceur-ws.org. [Online]. Available: http://ceur-ws.org/Vol-2380/paper_124.pdf.

[11] D. Gupta , S. Suman, and A. Ekbal, "Hierarchical Deep Multi-modal Network for Medical Visual Question Answering," 2020, arXiv.org. [Online]. Available: https://arxiv.org/abs/2009.12770.

[12] A. Lubna, S. Kalady and A. Lijiya, "MoBVQA: A Modality based Medical Image Visual Question Answering System," 2019, IEEE. [Online]. Available: https://ieeexplore.ieee.org/document/8929456.

[13] I. Allaouzi, B. Benamrou, M. Benamrou, and M. B. Ahmed, "Deep Neural Networks and Decision Tree classifier for Visual Question Answering in the medical domain," 2017(check), ResearchGate. [Online]. Available: http://ceur-ws.org/Vol-2125/paper_159.pdf

[14] Y. Liu, X. Zhang, F. Huang, L. Cheng, and Z. Li, "Adversarial learning with multi-modal attention for visual question answering," 2020, IEEE. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9174895.

[15] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, arXiv.org. [Online]. Available: http://arxiv.org/abs/1409.1556.

[16] 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.

[17] D. Kingma, and J. Ba, "ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION," 2015, arXiv.org. [Online]. Available: https://arxiv.org/abs/1412.6980.

[18] M. Malinowski, and M. Fritz, "A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input," 2015, arXiv.org. [Online].Available: https://arxiv.org/abs/1410.0210.

[19] 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.

[20] Lau, J.J., Gayen, S., Abacha, A.B., Demner-Fushman, D., 2018. A dataset of clinically generated visual questions and answers about radiology images. Scientific Data 5, 1–10.

[21] B. Liu, L. Zhan, L. Xu, L. Ma, et al, "SLAKE: A Semantically-Labeled Knowledge-Enhanced Dataset for Medical Visual Question Answering," 2021, ResearchGate. [Online]. https://ieeexplore.ieee.org/abstract/document/9434010.