

Medical Visual Question Answering

Nidhin Mohan
B180948CS

Computer Science and Engineering
National Institute of Technology, Calicut
Calicut, India
nidhin_b180948cs@nitc.ac.in

Abhinav B Naik
B170297CS

Computer Science and Engineering
National Institute of Technology, Calicut
Calicut, India
abhinav_b170297cs@nitc.ac.in

Shaaheen A M
B181134CS

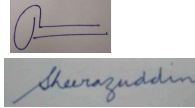
Computer Science and Engineering
National Institute of Technology, Calicut
Calicut, India
shaaheen_b181134cs@nitc.ac.in

Group No. 42

Guided by:

Dr. Saidalavi Kalady

Dr. S. Sheerazuddin



Abstract—The Visual Question Answering(VQA) domain explores the problem of answering questions regarding images. VQA in the medical domain(Med-VQA), has enormous potential and applications in the medical industry, such as assisting doctors in the field of radiology. Present Med-VQA frameworks employ deep learning approaches and image and question classification techniques. However, the quantity and diversity of questions and images in the medical domain, makes the task of Med-VQA challenging. Here we present our Med-VQA framework evaluated using the SLAKE dataset, which contains 642 images and 14,028 questions, which is larger compared to other datasets such as VQA-RAD which has 315 images and 3,515 questions. We use pretrained VGG16 model for obtaining image features. GloVe word embeddings and LSTMs are used for obtaining question features. The image and question features are then concatenated together, which is then passed to a classifier to obtain the answers. The currently available frameworks reach accuracies in the range of 50-70

Index Terms—Visual Question Answering, Med-VQA, SLAKE, VQA-RAD, CNN, GloVe, LSTM, Bilinear Attention Networks

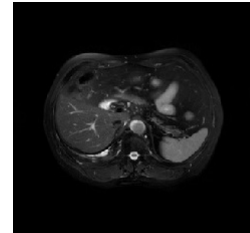
I. INTRODUCTION

Artificial Intelligence(AI) and Deep Learning has seen many advancements in the recent years, especially in the fields of computer vision and natural language processing. With advancements in these fields, more challenging problems are arising for which solutions are to be found. Many of those problems tend to combine image and language features, such as image captioning and visual question generation.

One such area of research is in the problem of Visual Question Answering(VQA). Visual Question Answering aims to correctly answer a question related to a given image. The task involves understanding both image and question features in order to give a correct answer. For example, given an image of a car, a person can ask "Which color is the car in the image?" or "What is shown in the image?". The VQA framework is expected to correctly answer the questions asked. VQA consists of both simple closed-ended questions like a question which answers either yes or no, or more complex open ended questions to which the answer can be anything. Certain questions also require better reasoning capabilities to

be given a correct answer. Research is still being done to develop a highly efficient and accurate VQA models.

Visual Question Answering has many applications in the field of robotics, assisting visually impaired people, answering questions regarding complex images and in the field of medicine. In medical visual question answering, a complex medical image is given as the input to which doctors can ask any clinical question regarding the image, in order to get a second opinion about the diagnosis. It can also be used by doctors to detect tumours or abnormalities in a given X-Ray, CT or MRI.



Q: What modality is used to take the image?
A: MRI
Q: Does the above picture contain kidney?
A: No

Fig. 1. An example of Medical VQA where an image and question are given as input and the software must correctly answer the question. Here, an example each of open-ended and closed-ended questions are shown

Tackling the problem of Medical VQA has its difficulties, mainly when procuring the medical datasets to train the VQA model. The literature available for Medical VQA is quite limited and the questions need to be answered with much greater accuracy. Med-VQA questions include closed ended questions such as asking for the presence of a particular organ in the image or detecting the presence of an abnormality, to more complex open ended questions such as identifying the abnormality or organs present in the given image.

In this paper, we are currently aiming to answer questions through classification means. We first obtain the image features using Convolutional Neural Networks. Then we obtain the

question features using GloVe word embeddings [1] and Long Short Term Memory Networks(LSTM). We then concatenate the features together to obtain a single representation of the image-question pair, which is then given to a classifier which predicts the answers from a given set of answers.

II. LITERATURE SURVEY

In this section, we describe the recent works related to VQA and Medical VQA.

[3] describes the AIML team's contribution to 2020 ImageCLEF Medical Domain Visual Question Answering (VQA-Med) challenge. The team approached VQA problem using a knowledge inference methodology called Skeleton-based Sentence Mapping (SSM). The method summarized questions with similar sentence structures into a unified backbone. Due to this technique's help, they were able to solve the VQA task as a multi-task image classification problem. They also used a class-wise and task-wise normalization method to optimize multiple tasks in a single network. This enabled them to apply a multi-scale and multi-architecture ensemble strategy for a robust prediction. These methods helped them secure the first position in the challenge with a score of 0.496 in accuracy and 0.542 in BiLingual Evaluation Understudy(BLEU) [4] score.

[5] improves upon the reasoning part of a VQA framework. They treat the problem as a classification task. They introduce a Question Conditioned Reasoning(QCR) module and a Type Conditioned Reasoning(TCR) module, since the performance of VQA frameworks in open ended questions is relatively low compared to closed ended questions. The question conditioned reasoning module extract the abundant task information present in the questions. The type conditioned reasoning module classifies the questions into open-ended or closed-ended questions based on the emphasis given to specific words that differentiate between an open-ended or closed-ended question. The introduction of QCR and TCR modules showed a significant increase in the accuracy of open ended questions, from about a percentage accuracy of 49.2 without either modules to a percentage accuracy of 60.0 with both modules.

[6] followed a question and image classification approach to tackle the VQA problem. The dataset used was generated from the MedPix database and consisted of 3,200 medical images and 12,792 Question-Answer(QA) pairs as training data. It was observed that the category of question can be determined from the question words, and hence they classified the questions into four categories- plane, modality, organ system and abnormality type questions. Based on the output of question classification, the image was classified using the four image classification models used, and the output given as the answer. The modality classification model was more complex than the other three models as certain modality classes had various subclasses. The framework gave a percentage accuracy of 75.2 for organ system type questions, 77.6 for plane type questions, 72 for modality type questions and 18.4 for abnormality type questions.

[7] followed a similar approach to [6]. They used a dataset of 4000 training images, where each image is associated with a question-answer pair. The questions were of two types- closed-ended questions that asked whether the given image is normal or abnormal, and open-ended questions that asked which abnormality was present in the given image. The question was only used for determining the question type, closed or open-ended, and thus it was overall treated as an image classification task. Image features were extracted using pretrained VGG16 models with the final softmax layer removed and every layer except the last four frozen. They used two models, one for closed-ended questions and another for open-ended questions. The model gave a best accuracy of 0.48 and a BLEU score of 0.511.

[8] follows a novel approach for the task of VQA on medical dataset with a model that generates answers in a sequence of words for a medical image-question pair. Here a model was made to modify and combine both Machine translation techniques and Image captioning. The dataset used by them was ImageCLEF MedVQA dataset provided by ImageCLEF as a part of the 2018 VQA-med challenge. Here they have built a deep neural network using CNNs and Image embeddings are obtained and have used GRUs (Gated Recurrent Units) for both encoder and decoder models. These helped them secure the first position in the challenge with a BLEU score of 0.188.

[9] proposed model follows a hierarchical multi-modal approach to tackle the VQA problem in the medical domain, where the top-level task is an SVM based question segregation and the next-level task is answer prediction. Inception-Resnet is used to encode image feature and Bi-LSTM for question feature creation. The proposed model with a question segregation module, segregates the learning path based on the question types(Yes/No and others). They have evaluated their model on the RAD and CLEF18 medical VQA datasets. For CLEF+RAD dataset, BLEU score of 0.0365 was obtained. By comparing the proposed HQS-VQA technique with baseline models, a significant change in accuracy is recorded which shows effectiveness in their model.

[10] proposes a model called CGMVQA (Classification and Generative Model for Medical VQA). The authors' motive was to use the CGMVQA model to answer questions containing various medical images by transforming the strong artificial intelligence problem into multiple weak artificial intelligence problems. The dataset they used was the ImageCLEF 2019 VQA-Med dataset. They divided the questions into five different categories: yes-no, modality, plane, organ system and abnormality. They used methods like data augmentation on images and word tokenization on words to improve the features. Using ResNet152, they extracted the image features from different convolutional layers. They added three kinds of embeddings to express different word pieces in the text feature extraction part. Lastly they used a multi-head self-attention transformer to deal with sequence problems and reduced its parameters to cut the computational cost down. They adjusted the masking and output layers to change the functionality of the model. With this setup, the team was able to get a score

of 0.64 in accuracy, 0.659 in BLEU and 0.678 in WBSS.

[11] proposes a solution inspired by self-supervised pre-training of Transformer-style architectures for NLP, Vision and Language tasks. The team first pretrained their Multimodal Medical BERT(MMBERT) model on a VQA Dataset with Masked Vision-Language Modeling task. They used the ROCO dataset for this task. Then they loaded the model with weights from pretraining and fine tuned it further on the train split of respective two medical VQA datasets, i.e, VQA RAD and VQA Med 2019 datasets. For Image feature extraction, they used Resnet152 and extracted features from different convolutional layers. They experimented with 3 different settings for MMBERT, a general one, one with fine tuning for different question categories in the datasets, and a non-pretrained one. The second one achieved their best results with an accuracy of 0.672 and a BLEU score of 0.69.

We also looked into various meta learning models for training models to learn meta weights which can be used to learn weights of medical images using simple convolutional neural networks. [13] proposes a method of meta learning in which the parameters of a base model is updated using gradient descent from various adapted parameters and loss functions, in order to initialise the model parameters to learn according to a given task. However, it relies heavily on meta-annotation which is a limitation for the method.

[15] tries to overcome the problem of transfer learning weights for medical images and the limitations laid by MAML [13] by proposing a multiple meta-model quantifying process. In [15] multiple meta models are trained using MAML [13] and a list of candidate meta models is selected from the set of meta models based on their performance and difference of features from other meta models.

The overall framework of VQA used in these frameworks is similar- image feature extraction, question feature extraction, feature fusion and classification. Some of the papers have adopted simpler methods for modelling VQA frameworks. Other publications have added more modules for improving accuracy in areas like image and question feature extraction. On an average, the VQA models have an accuracy of around 0.6-0.7 and accuracy is shown more in closed-ended questions than in open-ended questions.

III. DATASET USED

We have used the Semantically-Labelled Knowledge-Enhanced(SLAKE) Dataset [12] for training our VQA models. On analysis of the Slake dataset, we have found that it contains 9,835 training examples and 2,094 testing examples, each example being a question-answer pair.

Total number of unique questions in training set is 1189 and total number of unique questions in the test set is 605. The questions were of 2 languages- English and Chinese. For our VQA model, we are only considering the English language questions.

The answers to the questions were also of various types. The questions were of multiple choice questions where the options are given in the question itself, yes/no questions, open-ended

TABLE I
RESULTS OF THE REVIEWED PAPERS SUMMARISED

Paper	Results
[3]AIML at VQA-Med 2020: Knowledge Inference via a Skeleton-based Sentence Mapping Approach for Medical Domain Visual Question Answering	Accuracy-0.496 BLEU-0.542
[5]Medical Visual Question Answering via Conditional Reasoning	Accuracy-60%
[6]Visual Question Answering in the medical domain based on deep learning approaches: A comprehensive study	Organ system type-75.2%, Plane type-77.6%, Modality type-72%, Abnormality type-18.4%
[7]The Inception Team at VQA-Med 2020: Pretrained VGG with Data Augmentation for Medical VQA and VQG	Accuracy-0.48, BLEU-0.511
[8]A Sequence-to-Sequence Model Approach for ImageCLEF 2018 Medical Domain Visual Question Answering	BLEU-0.188, WBSS-0.209
[9]Hierarchical deep multi-modal network for medical visual question answering	BLEU-0.0365, WBSS-0.173
[10]CGMVQA: A New Classification and Generative Model for Medical Visual Question Answering	Accuracy-0.64, BLEU-0.659, WBSS-0.678
[11]MMBERT: Multimodal BERT Pretraining for Improved Medical VQA	Accuracy-0.672, BLEU-0.69
[15]Multiple Meta-model Quantifying for Medical Visual Question Answering	PathVQA Free-Form Accuracy-0.134, Closed-Ended Accuracy-0.84, Overall-48.8 VQA-RAD Open-Ended Accuracy-53.7, Closed-Ended Accuracy-75.8, Overall-67.0

wh-questions with only a single answer and open-ended wh-questions with more than one answer. For our current project, we are only considering questions with a single answer-multiple choice, yes/no or open-ended.

TABLE II
QUESTION LANGUAGES

Language	Training	Test
English	4,919	1,061
Chinese	4,916	1,033

Questions were of two base types- vqa and kvqa.

TABLE III
QUESTION BASE TYPES

Base Type	Training	Test
vqa	8,661	1,827
kvqa	1,174	267

Questions were divided into 10 types based on their content. They are- Color, Abnormality, Modality, KG, Size, Organ, Plane, Shape, Position, Quantity.

TABLE IV
QUESTION CONTENT TYPES

Content Type	Training	Test
Color	268	65
Abnormality	1,410	311
Modality	1,070	228
KG	1,174	267
Size	650	124
Organ	2,543	507
Plane	505	116
Shape	83	12
Position	1,695	375
Quantity	437	89

We also analysed the starting words of each questions. Questions were either of the form:

- 1) Starting with is/does/are/in/do/can
- 2) Starting with what/where/how/why/when/which

TABLE V
QUESTION STARTING WORDS

Starting Words	Training	Test
Is/are/does/can/do/in	1,726	369
What/where/why/who/when/ho	3,193	692

There were a total of 642 images in the dataset. There were 586 images in the training set and 180 images in the test set.

The images were of 3 different modalities- X-Ray, CT or MRI.

TABLE VI
IMAGE MODALITY TYPES

Modality Type	Training	Test
X-Ray	2,808	740
CT	4,819	865
MRI	2,208	489

The location of the parts were of 10 area- Chest_heart, Abdomen, Chest_lung, Lung, Neck, Chest_mediastinal, Pelvic Cavity, Brain_Tissue, Brain_Face and Brain.

TABLE VII
IMAGE LOCATIONS

Location	Training	Test
Chest_heart	187	0
Abdomen	3,041	607
Chest_lung	283	21
Lung	3,406	828
Neck	264	75
Chest_mediastinal	33	21
Pelvic Cavity	434	83
Brain_Tissue	1,394	322
Brain_Face	250	60
Brain	543	77

There were a total of 484 unique training set answers and 270 unique test set answers. Answers were of two types- closed-ended and open-ended.

TABLE VIII
ANSWER TYPES

Type	Training	Test
CLOSED	3,881	836
OPEN	5,954	1,258

IV. DESIGN

In our VQA framework shown in Figure 2, there are mainly three sections- feature extraction, feature fusion and classification.

Feature extraction aims to extract the features from both the image and question for answer prediction. It can further be divided into image feature extraction and question feature extraction.

A. Image Feature Extraction

Image features are extracted using VGG16 [14]. A Convolutional Neural Network(CNN) is an algorithm which can be used to extract specific features regarding an image and learn the necessary weights which helps to distinguish images from one another. A CNN can be used to capture the relevant features in the image which could give a much better accuracy of answering to the VQA model.

VGG16 model takes images of fixed size input, 224x224x3, 3 being the number of channels. We used a VGG16 model that had been pretrained on the ImageNet dataset [16]. The full VGG16 model consists of three Dense layers and one Softmax layer at the end, which outputs a 1000-valued vector. For our model, we do not include the final Dense and Softmax layers, and hence VGG16 outputs the extracted features of the image in a 7x7x512 array.

For our model, we load the images from file and pass through the VGG model to obtain the features of the images. Each image feature matrix is of size 7x7x512. This matrix is flattened into a 1x25088 array. Each image feature is combined into a single array with number of rows being the number of images, and each row being the image features obtained using VGG16.

Since the size of each image feature is extremely large, the dimensions of the image has to be reduced. For that, Principal Component Analysis(PCA) is performed. PCA is an effective linear algebraic method that is used for reducing the dimensionality of large datasets. In PCA, a standardized dataset of large dimensions is taken. The covariance matrix is then computed for the dataset, using which the eigenvectors and eigenvalues are computed. Principal Components are calculated using the eigenvectors. The data is finally projected along the principal components so as to reduce the dimensions, but at the same time preserve the necessary information. Using PCA, the dimensionality of our dataset has been reduced from 25,088 to 512. Thus for each image, a 512-dimension vector is computed to be passed for feature fusion.

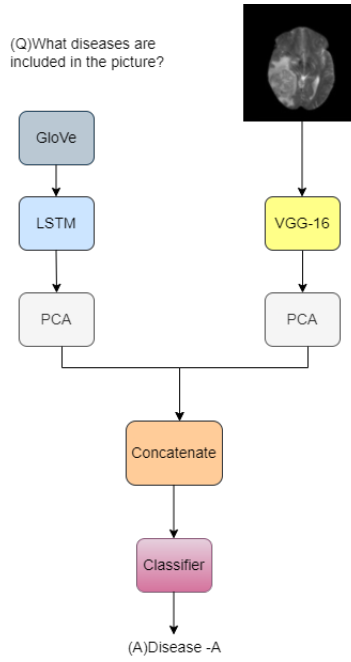


Fig. 2. Our Proposed Med-VQA Framework

B. Question Feature Extraction

Question features are extracted using Recurrent Neural Networks(RNNs) like LSTMs. Input to the RNN is a padded sequence of word embeddings created using GloVe word embeddings.

GloVe word embeddings provide a vector representation for each word in the question. The Euclidean distance between each vector is a good measure for the semantic similarity between two words.

Long Short Term Memory Networks have a feature of persistent memory, which enables them to learn the relevant features necessary and hence the features towards the end of the sequences can be learnt more accurately. This enables them to make much more effective predictions when compared to a simple Recurrent Neural Network.

We pass the word embeddings through the LSTM model to obtain Question Vectors similar to the image vectors we got earlier. The LSTM model was pretrained using a classification problem of the questions. For each question, the LSTM model outputs a vector of 800 features. All the question feature vectors are joined to create a matrix with number of rows equal to the number of questions for each dataset- training, testing and validation.

Principle Component Analysis is performed on the question features as well to reduce the dimensions from 800 to 512, similar to the images.

C. Feature Fusion

In our model, we have performed the simplest feature fusion method, that is concatenating the two features together. For each question, the corresponding image feature vector is

retrieved and concatenated to obtain a 1×1024 feature vector. Such a feature vector is obtained for each image-question pair and passed to the classification model.

D. Classification

The concatenated features are then passed onto a simple Classifier model consisting of two dense layers, with 512 and 256 units, and a softmax layer of 133 units, 133 being the total number of unique answers across the dataset.

V. IMPLEMENTATION PROGRESS

We have made a sample VQA model. We have implemented a VGG model for obtaining the image features. Since the training, testing and validation images are available in the same dataset, we have stored the features of each image in a separate file after performing PCA, and retrieve the vectors to avoid processing all images once more.

The questions are passed through the embeddings and the LSTM model, for all of training, test and validation data. We have also stored the question vectors after performing PCA, since the LSTM model is a pretrained model for the questions.

The image features of each question is obtained from the information regarding training, testing and validation data and then concatenated to obtain the fused vector.

The classifier has also been implemented and the fused features have been used to train the classifier to obtain the outputs. The classifier has been optimized using Adam optimizer, categorical crossentropy loss has been calculated and the accuracies noted. The classifier is then trained with a batch size of 64 for 20 epochs.

VI. RESULTS

Our model has produced less than satisfactory results on the training, testing and validation datasets.

The results shown are for all the image-question pairs in our datasets.

TABLE IX
ACCURACY RESULTS

Type	Accuracy
Training	19.49%
Validation	17.76%
Testing	18.16%

TABLE X
LOSS RESULTS

Type	Loss
Training	3.597
Validation	3.348
Testing	2.963

The low accuracies in the training, test and validation data can be due to the use of very simple components, particularly for feature fusion, extracting question features and for classification.

VII. CONCLUSION AND FUTURE WORK

The use of Medical VQA frameworks can help ease the process of diagnosis by providing additional knowledge to medical practitioners when examining X-Ray, CT or MRI images. Another option can be given to help in the diagnosis and for more accurate detection of tumors and abnormalities.

For our model, the low accuracies can be explained by our use of basic models for question feature extraction, feature fusion and classification.

Firstly, the use of concatenation to fuse our models can lead to very low accuracies compared to effective methods such as Bilinear Attention Networks [2]. Simply concatenating the two sets of features does not account for all the relevant image-question pairs necessary to answer the question. Hence, the feature fusion part has to be updated with Attention Networks.

The question feature extraction model can be substituted with a more effective pretrained model instead of making one of our own. Our question feature model wasn't trained properly, and the use of a pretrained LSTM model can lead to better accuracies.

Instead of a simple 3-layer classifier to predict the answers, a Multilayer Perceptron Classifier can be used to obtain more accurate answers.

Furthermore, the use of VGG16 can be substituted with other image feature extraction models, particularly ones that have been trained on medical images.

REFERENCES

- [1] Jeffrey Pennington, Richard Socher and Christopher D. Manning. "GloVe: Global Vectors for Word Representation". In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).
- [2] Jin-Hwa Kim, Jaehyun Jun and Byoung-Tak Zhang. "Bilinear Attention Networks". In arXiv:1805.07932.
- [3] Zhibin Liao, Qi Wu, Chunhua Shen, Anton van den Hengel and Johan W. Verjans. "AIML at VQA-Med 2020: Knowledge Inference via a Skeleton-based Sentence Mapping Approach for Medical Domain Visual Question Answering." CLEF (2020).
- [4] Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. "BLEU: a Method for Automatic Evaluation of Machine Translation". In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002.
- [5] Li-Ming Zhan, Bo Liu, Lu Fan, Jiaxin Chen, and Xiao-Ming Wu. 2020. Medical Visual Question Answering via Conditional Reasoning. In Proceedings of the 28th ACM International Conference on Multimedia (MM '20). Association for Computing Machinery, New York, NY, USA, 2345–2354. DOI:https://doi.org/10.1145/3394171.3413761
- [6] Aisha Al-Sadi, Mahmoud Al-Ayyoub, Yaser Jararweh, and Fumie Costen. "Visual question answering in the medical domain based on deep learning approaches: A comprehensive study", Pattern Recognition Letters, vol. 150, pp. 57–75, 2021. doi:10.1016/j.patrec.2021.07.002.
- [7] Aisha Al-Sadi, Hana AL-Theibat, Mahmoud Al-Ayyoub. (2020). The Inception Team at VQA-Med 2020: Pretrained VGG with Data Augmentation for Medical VQA and VQG.
- [8] R. Ambati and C. Reddy Dudyala. "A Sequence-to-Sequence Model Approach for ImageCLEF 2018 Medical Domain Visual Question Answering," 2018 15th IEEE India Council International Conference (INDICON), 2018, pp. 1-6, doi: 10.1109/INDICON45594.2018.8987108.
- [9] Deepak Gupta, Swati Suman, Asif Ekbal. (2020). Hierarchical Deep Multi-modal Network for Medical Visual Question Answering.
- [10] F. Ren and Y. Zhou. "CGMVQA: A New Classification and Generative Model for Medical Visual Question Answering," in IEEE Access, vol. 8, pp. 50626-50636, 2020, doi: 10.1109/ACCESS.2020.2980024.
- [11] Y. Khare, V. Bagal, M. Mathew, A. Devi, U. D. Priyakumar and C. Jawahar. "MMBERT: Multimodal BERT Pretraining for Improved Medical VQA," 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), 2021, pp. 1033-1036, doi: 10.1109/ISBI48211.2021.9434063.
- [12] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, Xiao-Ming Wu. "SLAKE: A Semantically-Labelled Knowledge-Enhanced Dataset For Medical Visual Question Answering". In:arXiv:2102.09542v1(18 Feb 2021).
- [13] Chelsea Finn, Pieter Abbeel, Sergey Levine. "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks". In:arXiv:2106.15367(19 Jun 2021).
- [14] Karen Simonyan, Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In:arXiv:1409.1556(4 Sep 2014).
- [15] Tuong Do, Binh X. Nguyen, Erman Tjiputra, Minh Tran, Quang D. Tran, Anh Nguyen. "Multiple Meta-model Quantifying for Medical-Visual Question Answering". In:arXiv:2105.08913(19 May 2021).
- [16] J. Deng, W. Dong, R. Socher, L. -J. Li, Kai Li and Li Fei-Fei. "ImageNet: A large-scale hierarchical image database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, doi: 10.1109/CVPR.2009.5206848.



Abhinav B Naik



Shaaheen A M