IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 43, No. 1, February, 2026, pp. 441-450
Journal homepage: http://iieta.org/journals/ts

A Whip Smart Visual Question Answering System Using Deep Learning

Puviarasi Gowrinathan*, Valliyammai Chinnaiah, Parkavi Govindarajan

Madras Institute of Technology, Anna University, Chennai 600044, India

Corresponding Author Email: puviarasig@gmail.com

https://doi.org/10.18280/ts.430131

Received: 10 February 2025
Revised: 22 August 2025
Accepted: 25 January 2026
Available online: 28 February 2026

Keywords:
Intelligent Question Answering, Convolutional Neural Networks, long short-term memory, fusion model

ABSTRACT

The human brain understands the image and processes the text-based question, and infers an answer. It is vital to create a humanoid system that will come up with a response to the question posed about the image. An Intelligent Question Answering (IQA) system that apes the humanoid system is proposed. An intelligent Question Answering system is a taxing multimodal learning activity as it entails the knowledge of both visual and textual modalities. The approaches that are used to process the images and questions play a vital role in the performance of the system. The Convolutional Neural Networks (CNN) is used to extract the significant features present in the image, and the long short-term memory (LSTM) is used to extract the textual content from the query. The proposed system uses a fusion of two deep learning models that combines the image features generated by the CNN model with encoded text features generated by LSTM networks to generate answers relevant to the question asked. The proposed fusion model performed better than the existing models, achieving an overall accuracy of 86% and an F1 score of 0.85 with a test-train split of 60-40 and an optimal combination of various hyperparameters.

1. INTRODUCTION

The image-based question answering is well-known for assisting people with visual impairments through natural language interactions and automatic querying of videos. It can also be used to respond to a question in accordance with a chart image. A computer system is considered to examine an image and respond to natural language questions about it. The questions can be open-ended or structured, completely random and include multiple computer vision sub-problems. The model must thoroughly comprehend both the query and the image characteristics for answering the questions. The response can be structured or open-ended, depending on the context of the question being asked. The processing of both the image and question simultaneously is a tedious procedure. The image-based question answering is an artificial intelligence (AI) research problem that spans multiple disciplines. The user needs a lot of observation to answer a question about a chart. In that case, IQA will effectively generate relevant answers to the question asked in a natural language to save time. The majority of extant image-based QA approaches are based on learning fine-grained question and image feature so that richer multimodal feature representations may be acquired. The image-based QA task has been effectively applied to the creation of attention mechanisms in the field of deep learning. The earlier attention models concentrate on specific areas of the image in order to correctly respond to the question "What is the name of the weapon hidden in the hand?". The existing papers proposed the visual attention approach, and it has since become an intrinsic feature of the VQA model, which demands fine-grained visual understanding. The visual attention and text attention can be learned by co-attention methodologies that focus on both the significant portions of the image and the keywords of the question. The initial co-attention approaches learned the coarse interactions but neglected the dense interactions between each question word and the subsequent image area, leading to the failure to determine the relationships between any question word and any image area. As a result, early co-attention strategies have significant limitations. Another method used in earlier approaches was image captioning. Initially the image captioning model will generate the appropriate description of an image. The system will infer an answer to the question based on the description provided. This co-attention mechanism might provide a wrong answer to the question asked. The proposed fusion model shall focus on every part of the image and overcome the existing issues in answering the question. The main objective of the model would be effectively generated relevant and most answers to the questions based on the information obtained from the image with an improved model performance, increased reliability and optimized resource optimization.

2. RELATED WORK

2.1 Question answering system with text

Yang et al. [1] proposed an Event-oriented Visual Question

Answering (E-VQA) dataset including free-form questions and answers for real-world event concepts. Wakchaure et al. [2] proposed a three-layer architecture for the field of answer selection in community question answering. The SemEval 2015 CQA dataset is used to implement the model. The three layers are Convolutional Neural Networks (CNN), long short-term memory (LSTM), and Conditional Random Field. The future work for this implementation is to use BiLSTM, which could possibly get better results.

Zhou et al. [3] proposed a comparison study of two models, such as CNN and LSTM, to build an automatic question answering system. The study is conducted on three different datasets. The accuracy of LSTM is higher than that of the CNN model. Thus, LSTM is more suitable for processing textual sequence data. Although CNN is standard for image-type data, applying it to textual data had good results.

Shuai and Zhang [4] developed a film-field-based question-answering system that comprises three major parts, namely knowledge graph construction, question preprocessing, and answer generation. The question preprocessing is carried out by a Naive Bayes classifier, which classifies the questions based on the identification of user intention in the question. The knowledge graph is built with the help of the Neo4j graph database.

Anhar et al. [5] introduced question classification in a QA system by using one of the deep learning models, BiLSTM. The reason behind using BiLSTM is that it does not depend on a certain sentence pattern and also it is capable of multi-class classification. The accuracy achieved on multiclass BiLSTM is 90%, which is higher than the basic LSTM model, and a loss of 31% is observed.

Wang and Wang [6] proposed a question answering system for a disease knowledge base. The architecture of the proposed model consists of three parts, namely text classification task, entity recognition module, and attribute classification module. In the text classification task, the BERT model is used. There are three layers, namely BiLSTM, CNN, and CRF layers, used in the Entity Recognition module. There are three parts, namely BiLSTM, dense, and softmax layers, used for attribute classification. The joint performance of these models achieved an accuracy of 82%.

2.2 Question answering system text with image

Chen et al. [7] proposed a framework that brings a fresh new way of viewing the interpretation of audio-visual scenes through both general and specific representations, as well as aggregating multi-modalities by prioritizing question-related information.

Bi et al. [8] proposed DenseCapBert, a revolutionary dense-caption-aware visual question-answering model, for better visual reasoning. Specifically, to strengthen the VQA models, dense descriptions for the photos and a multimodal interaction method are used to combine the dense captions, images, and questions into a single, cohesive framework.

Lobry et al. [9] used remote sensing data, and the system answers image-related questions. The CNN architecture is implemented for the visual part to extract information from a 2D image. The RNN architecture is used for the language part to retrieve features, and both image and text vectors are fused by point-wise multiplication, which creates an end-to-end model to be trained for answer prediction. Chen et al. [10] proposed Multimodal Encoder-Decoder Attention Networks (MEDAN). The MEDAN consists of repeated Multimodal Encoder-Decoder Attention (MEDA) layers that can access rich and reasonable question features and image features by relating keywords in the question with the most important object regions in the image. The MEDA layer has an Encoder module, which is a type of question-guided attention and self-attention of images, and a Decoder module, which is a type of question-guided attention and self-attention of images. The dataset used in this instance is the benchmark VQA-v2.

Gupta et al. [11] proposed a combined model for answering questions about images. The dataset is transformed by the preprocessing layer that processes questions and images. The Bidirectional LSTM layer, which works on word embeddings generated by GloVe, is used to retrieve temporal relationships among question words. The benchmark VQA dataset was used to create an EfficientDet-based image, which includes an extraction component for efficiently processing images. Liu et al. [12] suggested a VQA model with adversarial learning and bidirectional attention to tackle the VQA problem when provided with a dataset of textual questions and their corresponding images. The proposed model makes use of a question-oriented attention mechanism to perform feature fusion after extracting all of the image's features. These models have an oversight in that the proposed model is unable to successfully remove the image's irrelevant features.

Park et al. [13] proposed a model that consists of two models, the VQA module and the Sensitive Attribute Prediction (SAP) module. The VQA module predicts various types of answers, and the SAP module predicts only sensitive attributes using the same inputs. The proposed method validates VQA, GQA, and the proposed VQA Gender datasets through extensive experiments. Yu et al. [14] implemented an image model in this paper as Faster R-CNN to extract features, and the text model is word embeddings, which are used for question encoding. The relational reasoning and the attention mechanism are used to combine the image and text models to provide the final answer to the image-related question. In this paper, Yang et al. [15] proposed a model that develops a co-attention model with an end-to-end deep network architecture to co-learn the question features and image, and to reduce the candidate answer space, the question type is the concatenated version of the multi-modal joint representation. The novel network architecture offers a cohesive framework of VQA by combining the co-attention mechanism with the question type. The CAQT incorporates the co-attention mechanism and the type of question into a single and unified VQA model. VQA is the dataset.

The proposed system that was put forward by Wu et al. [16] has attributes that the CNN-based attribute prediction model predicts. The image captions are generated by the attribute-based captioning generation model. The attributes predicted and captions generated are combined with external knowledge mined out of a large-scale knowledge base and inputted into an LSTM to produce the answer to the question. The words that are highlighted indicate the information needed to answer the question. This approach is reflected in many recent successes.

Chowdhury et al. [17] proposed an image model using a VGG architecture, and its output is given to a PCA algorithm to reduce dimensionality. The question model used is LSTM. The proposed model achieves an accuracy of 37.19%. A VQA model was proposed by Huang et al. [18], using LSTM and VGG models. The compressed images constitute the dataset.

Cai et al. [19] proposed a model which integrates both a convolutional neural network and a bidirectional long short-

term memory network to do effective semantic analysis of question pairs to extract more effective text features. The proposed model is a combination of co-attention and attention mechanism to learn the characteristics of the input and existing questions jointly to get the interactive vector representation of the question pair. For training, the dataset contains 100,000 question-answer pairs.

Liu et al. [20] proposed a new conditional reasoning mechanism to acquire efficient reasoning abilities for many different Med-VQA tasks in an adaptive manner. Additionally, to pre-train a visual feature extractor to learn Med-VQA on a large amount of unlabeled radiological images, contrastive learning is used. Antol et al. [21] created the benchmark dataset open-ended VQA containing 0.25M images, 0.76M questions, and 10M answers.

The related works emphasize on implementing the visual question answering system that addresses free-form questions based on annotations, subtitles, and captions. The LSTM, CNN, and RNN are used majorly for creating models for extracting the text feature and image feature. The models fail because either information in visual frames, textual subtitles, and annotations are insufficient to answer all questions. The models don't explore the complex interactions between textual and visual features.

The proposed system performs effective semantic analysis on the question and image to get more useful features of the text and image features to answer the questions about the image through the fusion of ResNet50 and LSTM.

3. PROPOSED SYSTEM

The proposed Intelligent Question Answering system will infer an answer to the text-based question asked by the user by processing both the image and the question. The main aim is to build an effective deep learning model that will provide relevant and specific answers to the questions asked. The CNNs are used to extract the important features in the image and LSTM are used to extract the textual features in the question. The proposed model combines the image features generated by the CNN model with encoded text features generated by LSTM networks to generate answers relevant to the question asked. The user input consists of an image and a question associated with the image. The input image is fed to the Resnet50 model for the extraction of the features. Resnet50 is the convolutional neural network architecture that consists of 48 convolutional layers. All the important features are extracted from the last hidden layer of the Resnet50 model. Simultaneously the input question is tokenized and represented as vectors using GloVe. The vector representation of the question is then fed to the LSTM model. Both the image and question features are combined using pointwise multiplication for the final prediction of the answer. The GloVe embedding is used mainly because the dataset has a fixed vocabulary and the primary focus is on effectively capturing semantic relationship between words based on global context. A FastText can be used instead of GloVe for the dataset that contain morphologically rich text and out of vocabulary words. The pointwise multiplication is preferred over simple concatenation for fusing image and question features in the proposed because it creates a richer, more focused joint representation that captures semantic intersections between modalities. The Pointwise multiplication is also better than attention mechanisms as it is computationally an efficient way to integrate features compared to the more complex calculations that are involved in attention mechanisms that requires computing attention weights. The workflow of the proposed model is shown in Figure 1.

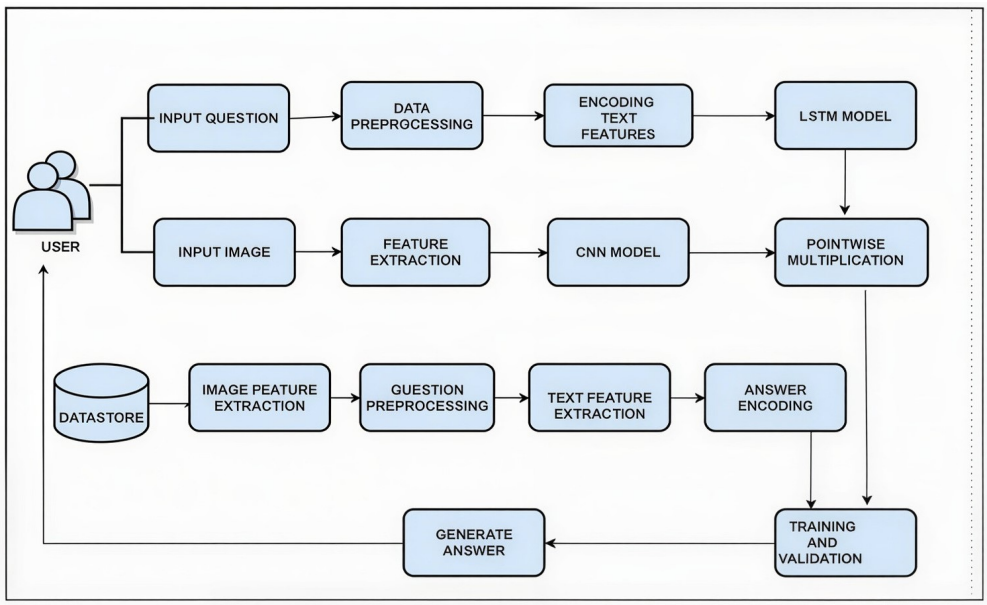

Figure 1. Visual question answering system

3.1 Data collection and preprocessing

The proposed model uses VQA. The dataset consists of three components, such as images, questions related to the image, and answers to the questions. The images are in png format. The questions and answers are stored in JSON format. Initially, all the three parts are combined together using a data frame for the further process. The text preprocessing is done on the question data. All the words in the question are converted to lowercase. The contractions are expanded, and then spaces on the leading and trailing portions of the question are removed. The sample preprocessed questions are given in Figure 2.

	im_path	ques	answ
0	/content/drive/MyDrive/vqa/sample_train/abstra...	what is the predominant color of the room?	brown
1	/content/drive/MyDrive/vqa/sample_train/abstra...	who will she give the bone to?	dog
2	/content/drive/MyDrive/vqa/sample_train/abstra...	what is the woman holding?	bone
3	/content/drive/MyDrive/vqa/sample_train/abstra...	how many people are in the room?	1
4	/content/drive/MyDrive/vqa/sample_train/abstra...	what color is the dog?	gray

Figure 2. Sample DataFrame

3.2 Image feature extraction

The input for the image model is images (.png format). The images are sorted in ascending order. Then the images are decoded and converted into float tensor values. The images are processed batchwise using the prefetch function, which is responsible for managing the next batch when the current batch is being processed. The Resnet50 model is trained on the batchwise images and the final 2 output layers are removed to extract the features alone in the input images. The features are stored in the form of an array in a numpy file. The image feature algorithm is given below:

Algorithm 1. Image feature extraction

input ← images
y = f eatures stored as array
Data frame ← images
while not at end of DataFrame do
  Read image
  Decode the image
  Convert into tensor values
  Resize the image into standard size of Resnet50 model
end while

3.3 Question processing and encoding

The questions are encoded using a GloVe embedding that produces a vector space with a meaningful substructure. The process of splitting a sentence into smaller units is called tokenization, which are words or otherwise known as tokens. The Tokenization is required before encoding the questions. A vocabulary array is created while performing tokenization. An integer is assigned to a vocabulary array with tokens as an index. The sentence is then converted into a list of numbers. The sequences are padded with zeros at the beginning to ensure the same length. The glove model is then loaded and creates an embedding matrix array which consists of a vector representation of words that are present in the vocabulary array of tokenization. The categorical variables are represented as binary vectors using one-hot encoding. It is used especially for the answer classes to encode the answer variable. The question feature algorithm is given below:

Algorithm 2. Question feature

input ← questions
y = questions stored as vector
Data frame ← questions
while not at end of DataFrame do
  Remove the leading and trailing spaces in the question
  Expand the contraction present in the question
  Tokenize the question
  Encode the questions using the glove to get the vector representation of question
end while

3.4 Model building and training

A new dataset is created with extracted image features along with questions and answers which can be directly given for training. The extracted features from the image which are stored in the form of a numpy array are given to the fully connected layer with Relu activation. At the same time, the vector representation of the question from the GloVe is fed to the LSTM model with two hidden layers followed by the fully connected layer with Relu activation. Both the image and question features are combined using pointwise multiplication. The flattening process is done followed by the fully connected layer with a softmax activation function which is used as a classifier. Then the fused model is fitted to the training dataset and evaluated on various metrics such as accuracy, precision, and recall. The hyperparameters such as learning rate, dropout, dense, batch size and epochs are fine tuned to achieve optimal resource utilization, increased model reliability and better performance metrics. The proposed model is tested with a sample image and question input.

Resnet50: The features are extracted using the Resnet50. The Resnet50 is a CNN architecture that uses CNN blocks multiple times. The challenge of training deepnet works can be addressed by the introduction of the residual blocks. The Transfer Learning model has a predefined shape. The predefined shape of Resnet50 is $1 \times 3 \times 224 \times 224$ where 1 is the batch size, 3 is a number of channels, 224 is the image width and the other 224 is the image height. There is a direct connection that skips some of the model's layers. The skip connection is known as the heart of residual blocks. The output is not the same because of this skip connection. The input is multiplied by the layer's weights and then the bias term is added without skip connection. Therefore, the input dimension may differ from the output dimension when a convolutional layer or pooling layers are used. The Batch normalization is done followed by each convolutional layer of the Resnet50 model. The skip connections of Resnet50 solve the vanishing gradient problem. Additionally, regularization will be used to skip those layers, if any layer affects the performance of the model. The Resnet50 architecture is shown in Figure 3.

Long Short Term Memory (LSTM): The text features of the question posed is extracted using LSTM. The LSTMs are artificial neural networks applied in artificial intelligence and deep learning. A LSTM model is also a recurrent neural network that can learn long-term dependencies of data. The model has a recurring module that has a combination of four layers interacting with one another. The cell state information in LSTM regulates the gate. The LSTM has two mechanisms namely the forget mechanism and the saving mechanism.

The forget mechanism forgets insignificant information that is not worth remembering. The saving mechanism saves important information which will be helpful in the future. Figure 4 represents the architecture of LSTM. The unit LSTM has the ability to remember information over time intervals of any length. A forget gate is responsible for removing information that is not important from the cell state. The architecture of LSTM model is shown in Figure 4.

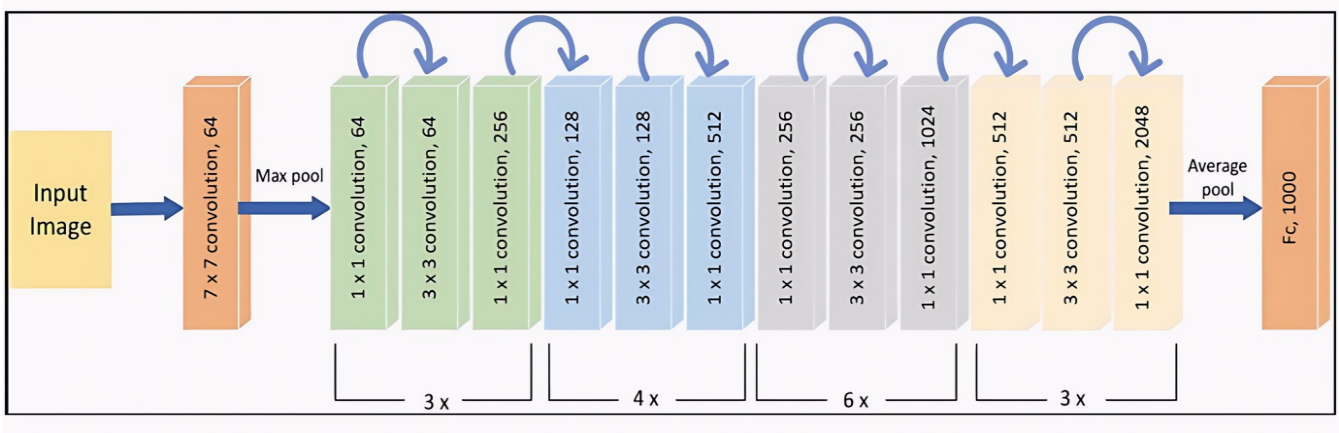

Figure 3. Resnet50 architecture [22]

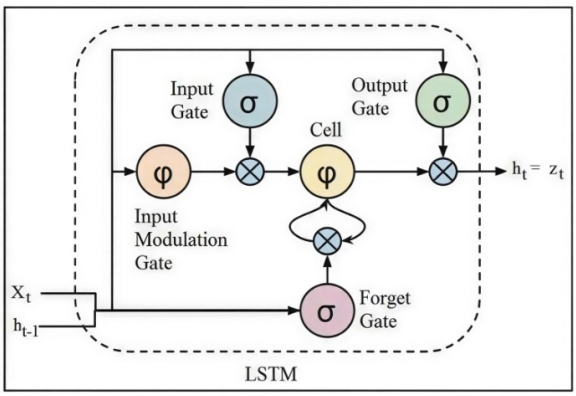

Figure 4. LSTM architecture [23]

4. PERFORMANCE AND RESULT ANALYSIS

4.1 Dataset

The dataset used is VQA v1.0 dataset obtained from Kaggle (https://www.kaggle.com/datasets/dmytruto/vqa-abstract-scenes). The benchmark dataset has three parts for training namely the image folder, questions file, and the answer file.

The questions and answers are stored in JSON format. The images are saved in .png format. A total of 6K images are used for training the model. There is a total of 15K questions present in the question file. The question file contains 3 fields namely question, question id, and image id. The answer file contains 5 fields namely image id, answer type, question id, question type, and answer. Most of the question length is 5 to 7 words and the maximum length of a question is 18 words. A very large proportion of the responses are yes/no, other than that the most common response consists of either colors or numbers.

4.2 Result analysis

Data analysis: The exploratory data analysis summarizes the main characteristics of data sets by analyzing them, often using visual methods. The count of words in the question is visualized as shown in Figure 5. Most of the question length is 5 to 7 words. The count of the top 20 answer classes is shown in Figure 6. The Boxplot representation for Question length is shown in Figure 7. The 50% of the questions contain 5 to 7 words. The majority of the questions are less than 10 words, with only few exceeding that limit.

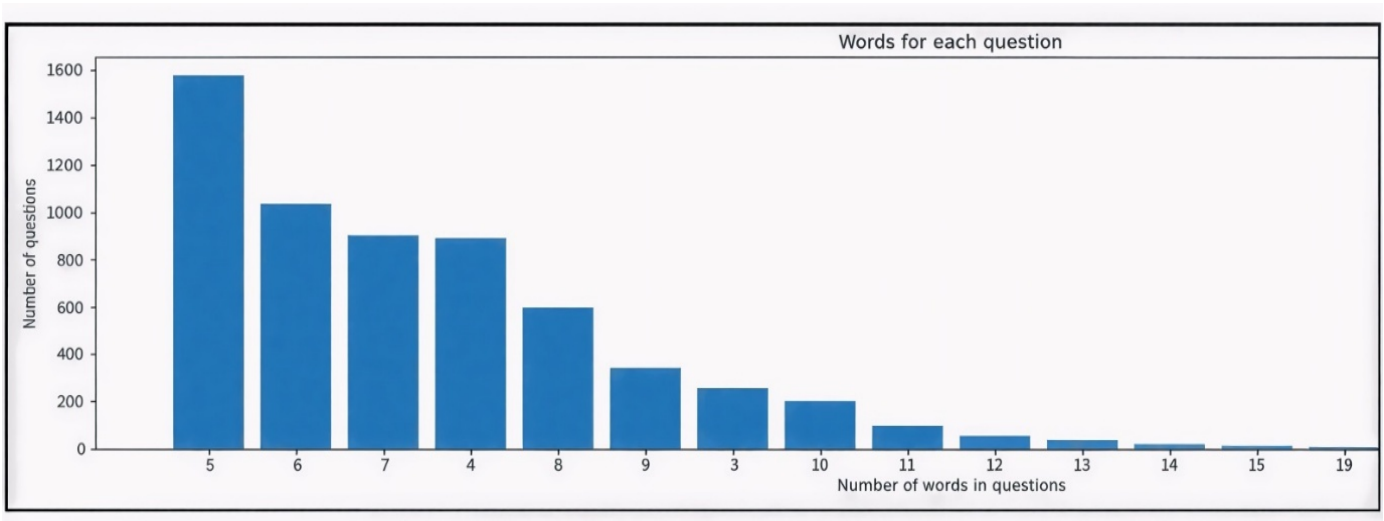

Figure 5. Word count of question

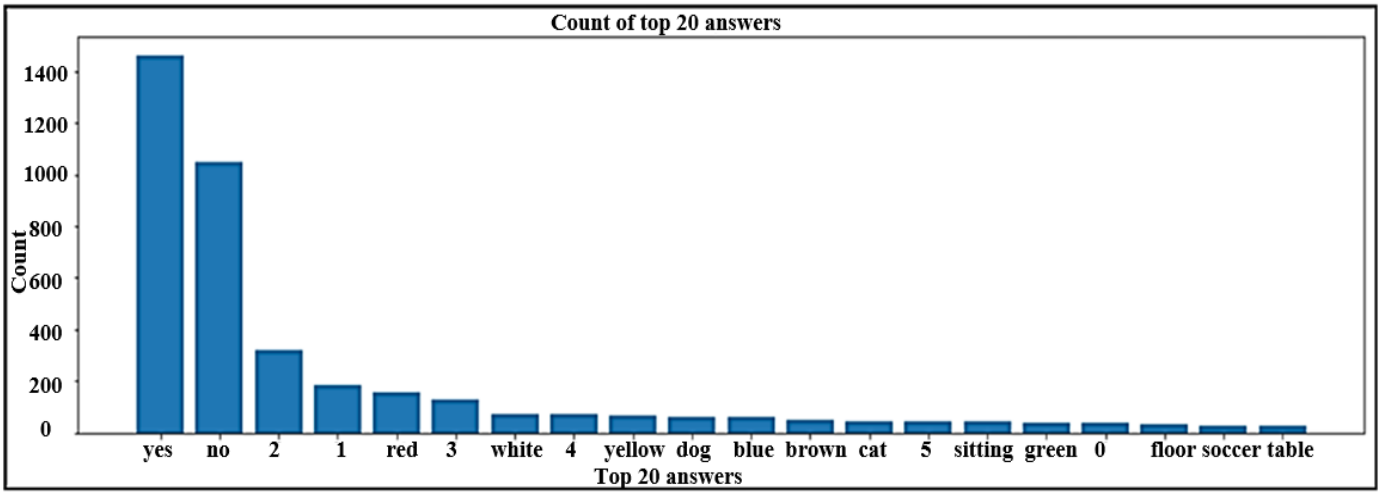

Figure 6. Count of top 20 answer classes

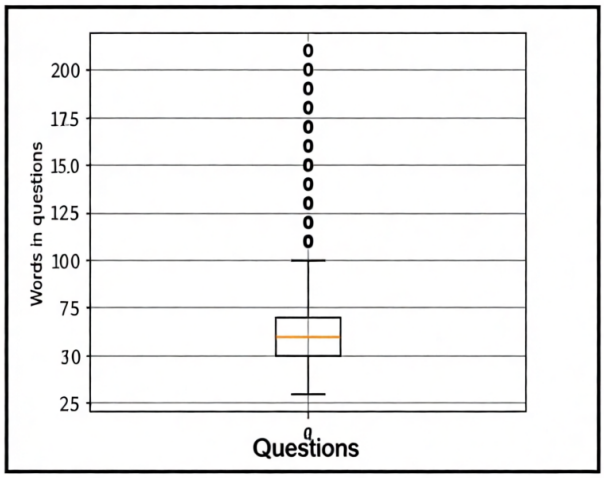

Figure 7. Boxplot for question length

4.3 Performance metrics

Accuracy: The accuracy is used to find the ratio of accurately predicted tuples to the total tuples.

Accuracy = (true positive+ true negative) / total tuples

Precision: The precision calculates the ratio of the number of true positives to the total number of predicted positives and is also called exactness.

Precision = true positive / (true positive + false positive

Recall: The recall calculates the ratio of true positives to total positives

Recall = true positive / (true positive + false negative)

F1-score: The F1-score is calculated as the weighted average of precision and recall

4.4 Training on various train-test split

Table 1. Training the dataset on various train-test split

No.	Train-Test Size	Accuracy	Precision	Recall
1	60-40	84.5	89	78.7
2	70-30	80.8	86.7	73.0
3	75-25	77.8	88.9	69.9
4	90-10	77.2	83.8	68.9

The dataset is trained on the proposed model with various train-test splits and performance metrics are listed in the following table. The split consists of 60% training and 40% testing obtained better performance than other splits on single runs. Table 1 shows the results with respect to various train-test split.

4.5 Model accuracy and loss

The accuracy of the various models in comparison with the proposed model is given in Figure 8. The loss of the various models in comparison with the proposed model is given in Figure 9. The proposed model accuracy is depicted in Figure 10 has higher accuracy than the existing models. The proposed model loss graph is depicted in Figure 11 has much lower loss than the existing models. The proposed IQA system has achieved 86% accuracy with the precision of 90%, recall of 80% and F1 score 0.85 with respect to number of epochs by fine tuning hyper parameters for the improved performance as shown in Figure 10. The proposed model loss graph plotted with the number of epochs is shown in Figure 11.

It shows the proposed model loss is gradually decreasing. The performance evaluation of various models such as provides Resnet50- LSTM, VGG16-LSTM, Mobilenet-LSTM, VGG19-LSTM, Densenet-LSTM, InceptionV3-LSTM, VGG16-BiLSTM, Resnet50-BiLSTM, InceptionV3-BiLSTM, VGG19-BiLSTM with respect to Precision, Recall, and F1-score is shown in Figure 12.

The proposed model's accuracy is compared with all other models, such VGG16-LSTM, Mobilenet-LSTM, VGG19-LSTM, Densenet-LSTM, InceptionV3-LSTM, VGG16-BiLSTM, Resnet50-BiLSTM, InceptionV3-BiLSTM, VGG19-BiLSTM as shown in Figure 13.

The models are trained on different train-test splits. The hyper parameters are tuned to find optimal combination for better results. The best results of all the models are used in the comparative study.

The proposed system provides the most relevant answers to the questions asked about the image and has achieved a better performance in terms of accuracy.The Figure14 depicts the answer to a given question based on the image by the question answering system.

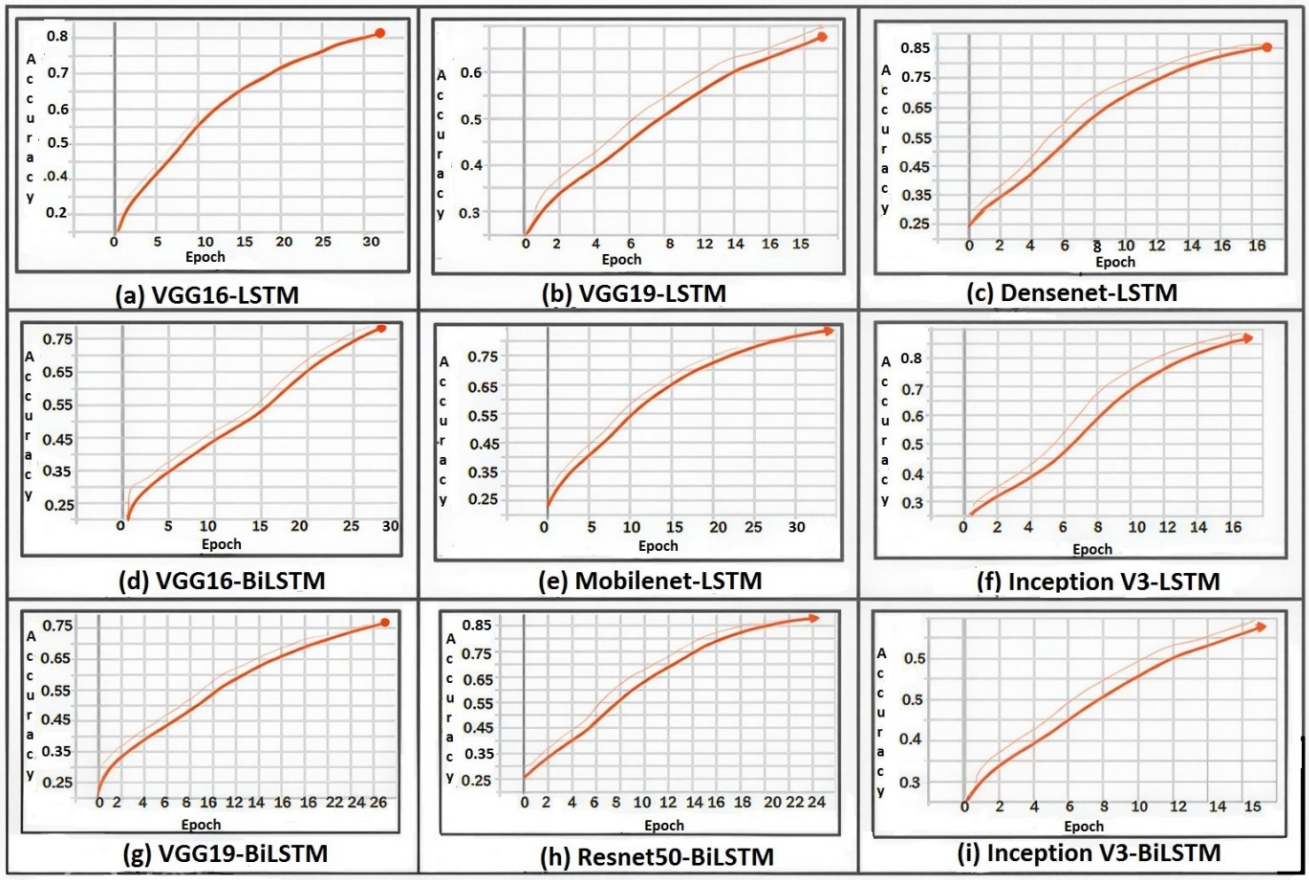

Figure 8. Accuracy graph of various models

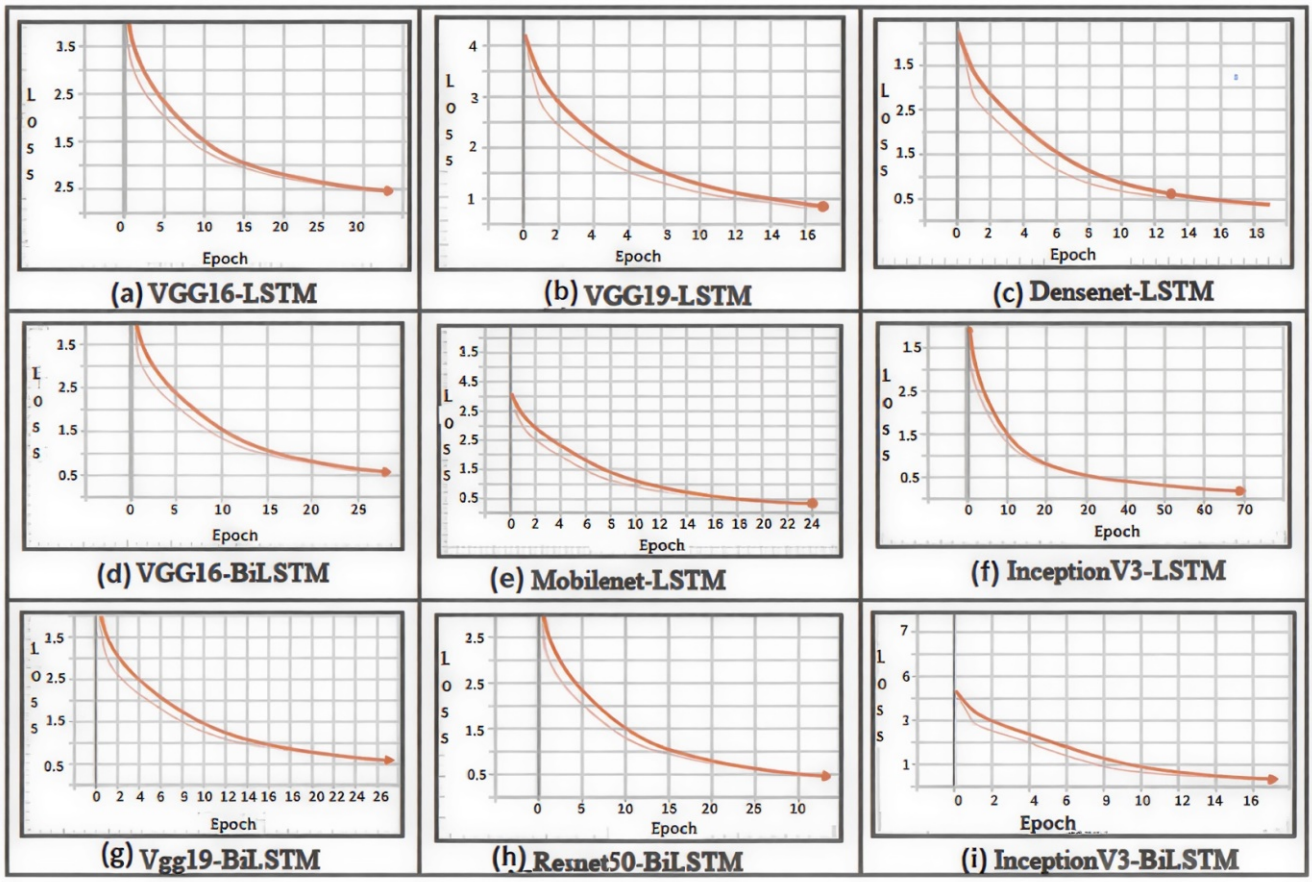

Figure 9. Loss graph of various models

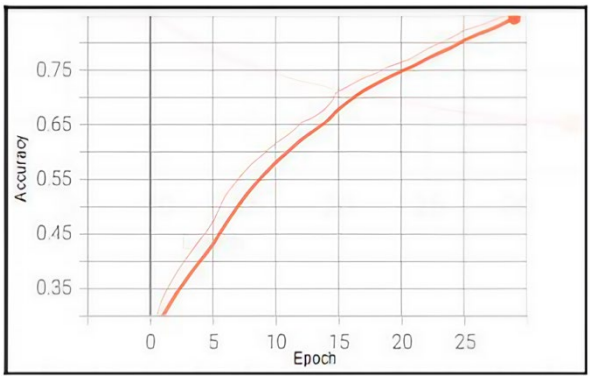

Figure 10. Proposed model accuracy

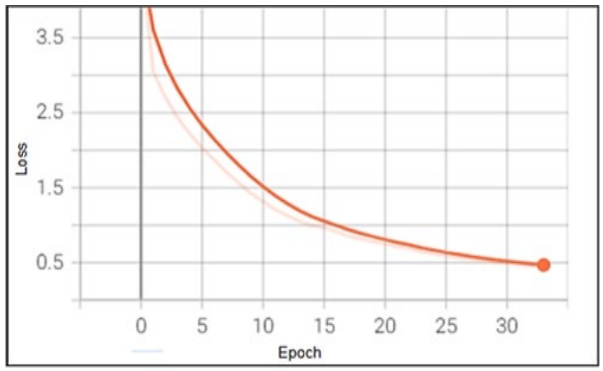

Figure 11. Proposed model loss

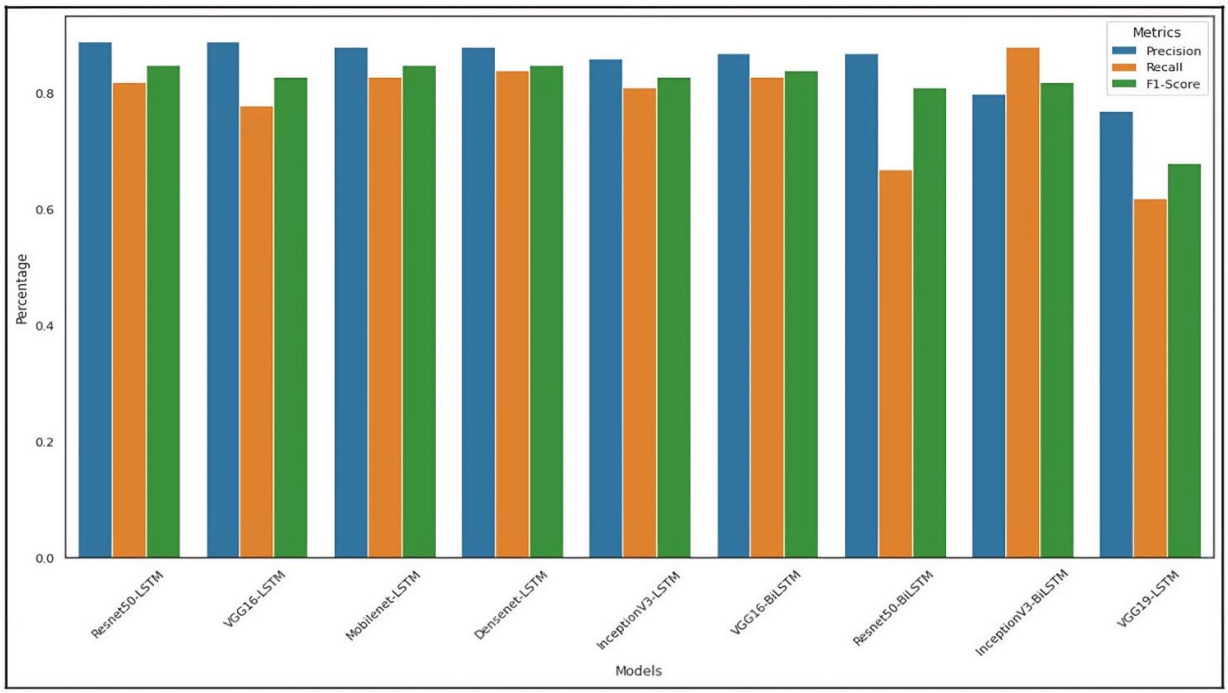

Figure 12. Performance of models

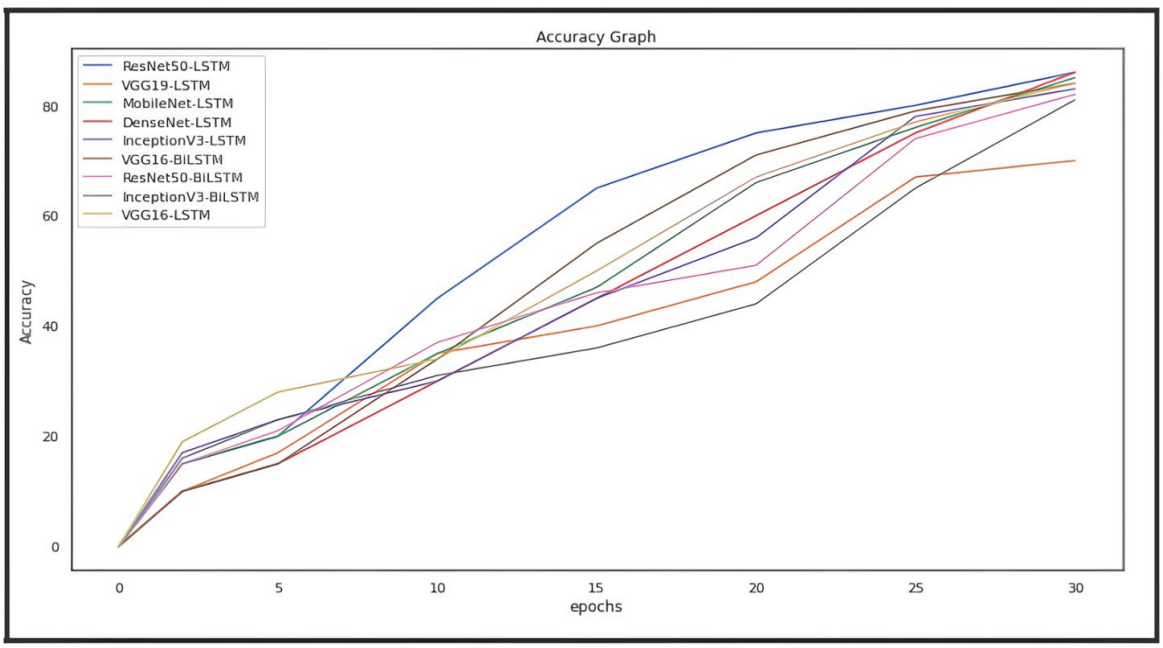

Figure 13. Accuracy comparison of models

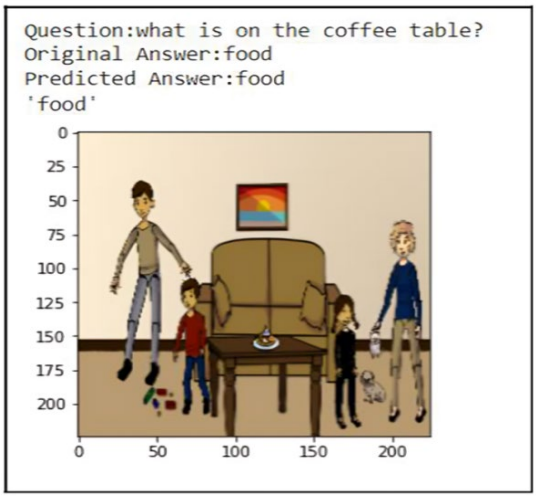

Figure 14. Sample result of prediction

5. CONCLUSION AND FUTURE WORK

A conclusion is implied by the human brain when it interprets the visual and reads the text-based question. It is crucial to develop a humanoid system that can produce answers to the questions regarding images. The proposed system mimics the humanoid system that implements deep learning approach to find a response to image related query.

A benchmark VQA dataset is used to train the proposed system. The CNNs are used to derive important features in the image and the LSTM are to derive textual features from the question. The image features generated by the Resnet50 model and the text features generated by LSTM networks are combined by the proposed system to generate relevant answers to the question asked. The proposed model's efficiency is evaluated in comparison to several deep learning models by testing with an image and asking questions about it for experimental purposes. The proposed model achieved an F1 score of 0.85 and an overall accuracy of 86%.

The experimental results show that proposed Resnet50-LSTM model outperforms the state-of-the-art techniques. The model is trained well owing to the fact that the predicted answer is matching with the actual answer. In the future, the proposed model can be used to analyze videos and better results can be achieved by fine-tuning the models. The system can be used to examine surveillance video to find out the hidden information present in the video for investigation purposes and this system can also be used for analyzing graph based images. This can be achieved by extracting textual features using BERT or BiLSTM with FasText embedding and attention mechanism integrated with resnet50 for the enhanced feature extraction from the visual frames of the videos.

REFERENCES

[1] Yang, Z., Xiang, J., You, J., Li, Q., Liu, W. (2023). Event-oriented visual question answering: The E-VQA dataset and benchmark. IEEE Transactions on Knowledge and Data Engineering, 35(10): 10210-10223. https://doi.org/10.1109/TKDE.2023.3267036

[2] Wakchaure, M., Kulkarni, P. (2019). A scheme of answer selection in community question answering using machine learning techniques. In 2019 International Conference on Intelligent Computing and Control Systems (ICCS), Madurai, India, pp. 879-883. https://doi.org/10.1109/ICCS45141.2019.9065834

[3] Zhou, X., Hu, B., Chen, Q., Wang, X. (2018). Recurrent convolutional neural network for answer selection in community question answering. Neurocomputing, 274: 8-18. https://doi.org/10.1016/j.neucom.2016.07.082

[4] Shuai, Q., Zhang, C. (2020). Question answering system based on knowledge graph of film culture. In 2020 International Conference on Culture-oriented Science & Technology (ICCST), Beijing, China, pp. 150-153. https://doi.org/10.1109/ICCST50977.2020.00035

[5] Anhar, R., Adji, T.B., Setiawan, N.A. (2019). Question classification on question-answer system using bidirectional-LSTM. In 2019 5th International Conference on Science and Technology (ICST), Yogyakarta, Indonesia, pp. 1-5. https://doi.org/10.1109/ICST47872.2019.9166190

[6] Wang, X., Wang, Z. (2020). Question answering system based on diease knowledge base. In 2020 IEEE 11th International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, pp. 351-354. https://doi.org/10.1109/ICSESS49938.2020.9237712

[7] Chen, Z., Wang, L., Wang, P., Gao, P. (2023). Question-aware global-local video understanding network for audio-visual question answering. IEEE Transactions on Circuits and Systems for Video Technology, 34(5): 4109-4119. https://doi.org/10.1109/TCSVT.2023.3318220

[8] Bi, Y., Jiang, H., Hu, Y., Sun, Y., Yin, B. (2023). See and learn more: Dense caption-aware representation for visual question answering. IEEE Transactions on Circuits and Systems for Video Technology, 34(2): 1135-1146. https://doi.org/10.1109/TCSVT.2023.3291379

[9] Lobry, S., Marcos, D., Murray, J., Tuia, D. (2020). RSVQA: Visual question answering for remote sensing data. IEEE Transactions on Geoscience and Remote Sensing, 58(12): 8555-8566. https://doi.org/10.1109/TGRS.2020.2988782

[10] Chen, C., Han, D., Wang, J. (2020). Multimodal encoder-decoder attention networks for visual question answering. IEEE Access, 8: 35662-35671. https://doi.org/10.1109/ACCESS.2020.2975093

[11] Gupta, R., Hooda, P., Chikkara, N.K. (2020). Natural language processing based visual question answering efficient: An EfficientDet approach. In 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, pp. 900-904. https://doi.org/10.1109/ICICCS48265.2020.9121068

[12] Liu, B., Zhan, L.M., Xu, L., Wu, X.M. (2022). Medical visual question answering via conditional reasoning and contrastive learning. IEEE Transactions on Medical Imaging, 42(5): 1532-1545. https://doi.org/10.1109/TMI.2022.3232411

[13] Park, S., Hwang, S., Hong, J., Byun, H. (2020). Fair-VQA: Fairness-aware visual question answering through sensitive attribute prediction. IEEE Access, 8: 215091-215099. https://doi.org/10.1109/ACCESS.2020.3041503

[14] Yu, J., Zhang, W., Lu, Y., Qin, Z., Hu, Y., Tan, J., Wu, Q. (2020). Reasoning on the relation: Enhancing visual representation for visual question answering and cross-modal retrieval. IEEE Transactions on Multimedia,

22(12): 3196-3209. https://doi.org/10.1109/TMM.2020.2972830
[15] Yang, C., Jiang, M., Jiang, B., Zhou, W., Li, K. (2019). Co-attention network with question type for visual question answering. IEEE Access, 7: 40771-40781. https://doi.org/10.1109/ACCESS.2019.2908035
[16] Wu, Q., Shen, C., Wang, P., Dick, A., Van Den Hengel, A. (2017). Image captioning and visual question answering based on attributes and external knowledge. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(6): 1367-1381. https://doi.org/10.1109/TPAMI.2017.2708709
[17] Chowdhury, I., Nguyen, K., Fookes, C., Sridharan, S. (2017). A cascaded long short-term memory (LSTM) driven generic visual question answering (VQA). In 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, pp. 1842-1846. https://doi.org/10.1109/ICIP.2017.8296600
[18] Huang, L.C., Kulkarni, K., Jha, A., Lohit, S., Jayasuriya, S., Turaga, P. (2018). CS-VQA: Visual question answering with compressively sensed images. In 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, pp. 1283-1287. https://doi.org/10.1109/ICIP.2018.8451445
[19] Cai, L. Q., Wei, M., Zhou, S.T., Yan, X. (2020). Intelligent question answering in restricted domains using deep learning and question pair matching. IEEE Access, 8: 32922-32934. https://doi.org/10.1109/ACCESS.2020.2973728
[20] Liu, B., Zhan, L.M., Xu, L., Wu, X.M. (2022). Medical visual question answering via conditional reasoning and contrastive learning. IEEE Transactions on Medical Imaging, 42(5): 1532-1545. https://doi.org/10.1109/TMI.2022.3232411
[21] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D. (2015). VQA: Visual question answering. In Proceedings of the IEEE international conference on computer vision, pp. 2425-2433.
[22] Talo, M. (2019). Convolutional neural networks for multi-class histopathology image classification. ArXiv, abs/1903.10035. https://arxiv.org/pdf/1903.10035.
[23] Long Short-Term Memory (LSTM): Concept**.** https://medium.com/@kangeugine/long-short-term-memory-lstm-concept-cb3283934359.