IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2320-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 7, July 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.569

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007025 |

Visual Question Answering through Memory Augmented Networks

Dr. G.Arun Sampaul Thomas [1], N.S.S.D.Bhargav [2], R.Snehith [3] ,T.Durga Prasad [4], V.Ganesh [5]

Associate Professor, Department of Computer Science and Engineering, JB Institute of Engineering and Technology, Moinabad, Telangana, India[1]

B.Tech. Student, Department of Computer Science and Engineering, JB Institute of Engineering and Technology, Moinabad, Telangana, India[2,3,4,5]

ABSTRACT: In recent years, advances have been made in the disciplines of computer vision, object detection, and natural language processing (NLP). NLP is used by Artificial Intelligence (AI) systems, such as question answering models, to give the computer with a complete capability. A system like this can respond to natural language questions about any part of an unstructured text. The objective of Visual Question Answering (VQA), which is to design a system that can answer natural language queries about images, can be accomplished by combining NLP and computer vision. For VQA, a number of systems based on deep-learning architectures and learning algorithms have been developed. Visual Question Answering (VQA) is a new issue domain in which multi-modal inputs must be processed in order to answer a natural language challenge. Visual and natural language processing, as well as abstract reasoning, are required by the solutions. This research proposes a VQA system that uses a deep convolutional neural network (CNN) to extract visual attributes and acquire deep knowledge from them. Our system achieves complex reasoning abilities and a natural language understanding, allowing it to accurately analyse the inquiry and provide an acceptable response. We emphasise on enabling the system to even answer the uncommon queries that are different from the ones used while training the model.

KEYWORDS: Visual Question Answering, convolutional neural network, multi-modal inputs.

I. INTRODUCTION

Visual Question Answering is a far more difficult task than image captioning because it usually requires information not visible in the image. This additional information could be anything from common sense to encyclopaedic expertise of a specific image aspect. VQA is a truly AI-complete task in this regard, as it necessitates multimodal knowledge beyond a single sub domain. This is why VQA is gaining popularity, since it serves as a proxy for assessing our progress toward AI systems capable of advanced reasoning, deep language understanding, and image interpretation.

Even when these answers are infrequent in the training set, memory-enhanced neural networks are utilised to predict accurate solutions to visual questions. The memory network integrates internal and external memory blocks, focusing on each training sample individually. Memory-enhanced neural networks have been shown in studies to have a long-term memory for unusual training samples. Because of the heavy-tailed distribution of answers in a typical VQA scenario, this is crucial for visual question answering. When a natural language question and a specific image is given, the task of the VQA system is to predict an answer that correctly fits the description of the given image

II. RELATED WORK

Kafle et al. propose a VQA system using a classical algorithm that uses Bayesian probability to predict the probability of the answer type given the input question and image. This answer type is then used to generate the actual answer. The model uses Bayes' probability rule to compute the probability. They use skipthought vectors to compute feature embeddings for questions. In 2016, Teney et al. proposed a VQA system that used graphs to represent the images and questions. A deep neural network was then built over this graph to identify the connection between the image and question from the graphical representation. The input image was encoded as a graph where each node in the graph

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007025 |

represented the objects in the image and the edges between these nodes represented the relative positions of those objects in the image. The question tokens were also encoded in a graph that represented the sequential connections and syntactic dependencies between the words in the question. Most of the other deep learning models proposed for VQA typically make use of the CNN/LSTM approach. CNNs are used to transform input images into their corresponding vector representations and extract meaningful insights from the images which can be used to train the model. One of the most popular models used for feature extraction from images is the VGGNet system proposed by Simonyan et al. Similarly pre-trained word embeddings are obtained to transform the question into its vector form. The most widely used word embeddings are Word2Vec and GloVe .These embeddings are then merged using different techniques to train the model on the combination of the image and question and then provide an answer.

III. METHODOLOGY

The suggested system is a CLEVR dataset-based Machine Learning model. The CLEVR dataset contains photographs of 3D-rendered objects, each with a collection of highly compositional questions divided into categories. The questions assess your knowledge of the numerous types of relationships that exist between these objects. The comprehension of the position relative to the other items is part of the relationships. To encode question words, we employ the GloVe (Global Vectors for Word Representation) word embedding model. The encoded words are fed into the LSTM one by one, resulting in a list of encoded output. ImageNet is a 22-layer convolutional neural network (CNN) that was pre-trained on the ImageNet dataset. Before jointly embedding the image and question features, we apply a co-attention technique to pay attention to the most relevant picture regions as well as textual terms. To improve our ability to remember rare question and answer pairs for visual question answering, we propose using memory-augmented networks. We employ the LSTM model in particular, which has an internal memory and an external memory that is controlled by LSTM. When training the model and predicting the answers to the questions, a softmax or sigmoid function is used. The system modules used for our proposed VQA system are:

- **Data Collection:** CLEVR provides a dataset which has about 1 lakh images and 8.6 lakh questions and answers for all training and validation sets. The biases found in previously used datasets are overcame with the emergence of CLEVR dataset.
- **Data Preprocessing:** This module involves question embedding which encodes the textual words of the question asked and normalization of the reference image.
- **Feature Extraction:** ImageNet is used for extraction of image features. CNN and LSTM are used for feature extraction. The layers in the CNN convolve over the input and produce feature maps for detecting data patterns.
- **Training the Model:** This phase involves training and testing the model. We use Logistic Regression for training. In particular, the softmax function is used to get the most probable answer as final prediction.

IV. EXPERIMENTAL RESULTS

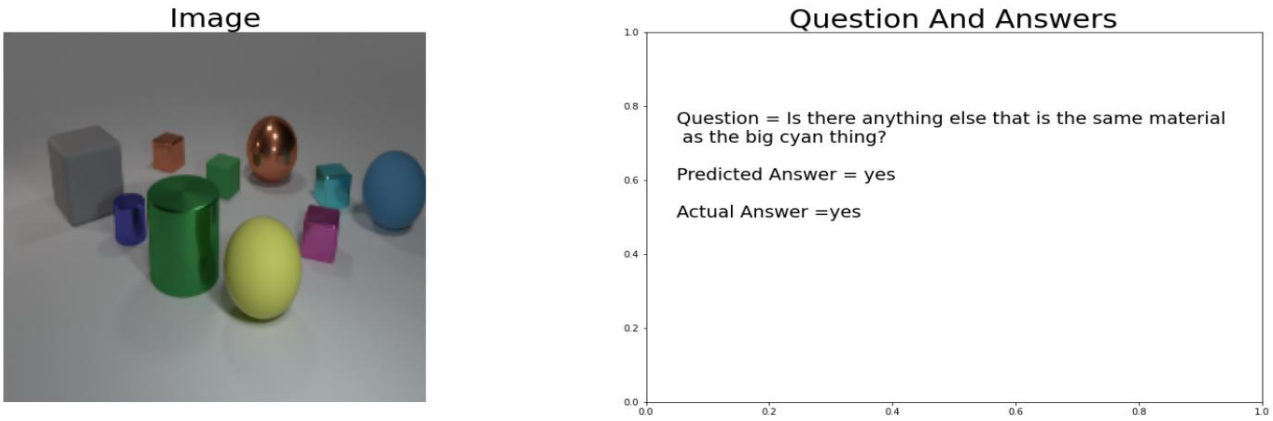

Fig.1. A test image from CLEVR dataset

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007025 |

When a question was asked about the relation between two objects i.e; two objects of same material but different colour in the image, the proposed system correctly predicted the answer.

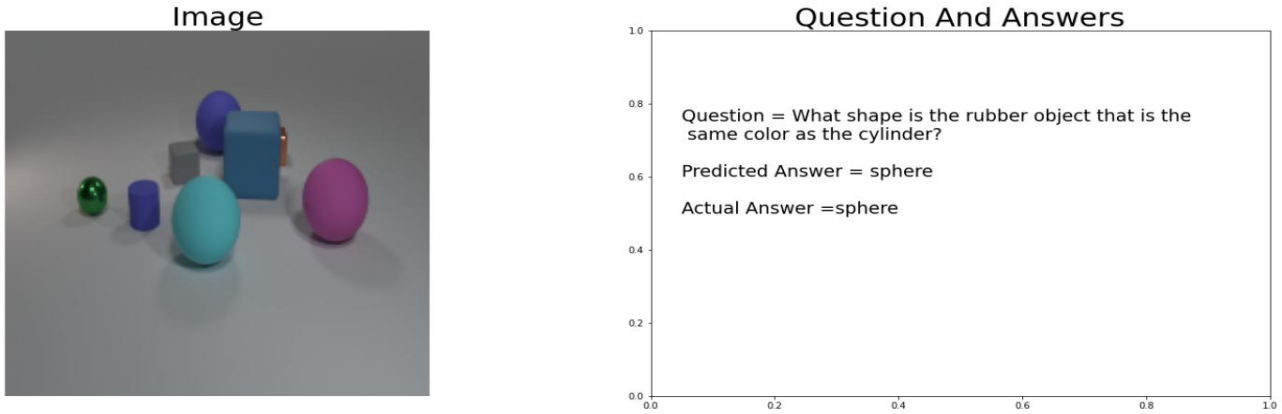

Fig.2. A test image from CLEVR dataset

When a question was asked about the shape of an object that is of a particular material but same colour as of a referenced object, the proposed system correctly predicted the answer

V. CONCLUSION

In visual question responding, we make a first attempt to clearly address the issue of rare notions. The proposed algorithm's core pipeline includes a co-attention module for selecting relevant picture regions and textual word characteristics, as well as a memory module for selectively paying attention to infrequent training data. The controller function of an LSTM module is to determine when to write or read from the external memory block. The features for learning a classifier that predicts replies are the outputs of the enhanced LSTM. On two large-scale benchmark datasets, the proposed approach outperforms existing VQA systems, and it is shown to successfully answer questions incorporating unusual concepts where other VQA methods fail.

REFERENCES

[1] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.
[2] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, “Supervised learning of universal sentence representations from natural language inference data,” arXiv preprint arXiv:1705.02364, 2017.
[3] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, “Bidirectional attention flow for machine comprehension,” arXiv preprint arXiv:1611.01603, 2016.
[4] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, “Vqa: Visual question answering,” in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2425–2433.
[5] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, “Visual7w: Grounded question answering in images,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4995–5004.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.569

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details