



Explainable Visual Question Answering Using Multimodal Deep Learning

Bodigiri Sai gopinadh

Assistant Professor, Department of Mathematics, GMR Institute of Technology Deemed to be University,
Rajam, India.

Abstract:

Visual Question Answering (VQA) integrates computer vision and natural language processing to answer questions about images. This paper proposes an explainable VQA framework that combines Convolutional Neural Networks (CNNs), language embeddings, and multimodal reasoning to improve answer accuracy and interpretability. The system extracts visual features from images, processes textual questions, and performs feature fusion for answer prediction. Experimental observations demonstrate the effectiveness of the framework for image understanding, accessibility applications, and intelligent human-computer interaction.

Keywords: Visual Question Answering, CNN, Deep Learning, Multimodal Learning, Explainable AI, NLP

1.INTRODUCTION

Visual Question Answering is a research area about building an AI system to answer questions presented in a natural language about an image. A system that solves this task demonstrates a more general understanding of images: it must be able to answer completely different questions about an image, oftentimes even addressing different sections of the image. For all the images, our AI system should be able to localize the subject being referenced to in the question and detect it and should have some common-sense knowledge to answer it. Visual Question Answering (VQA) allows people to ask natural language open-ended, multiple choice, and common-sense questions about the visual world. Visual data like images and videos are all around us and natural language is how we communicate so if we connect these two can lead to lots of applications. Like, visually impaired users do not have access to visual data but through this, they can access visual information. It also helps in conversations with future robots, and to analyze visual data faster for camera surveillance. We already have projects like image captioning and object detection but they can be very generic so we need a Visual Question Answering task where we can find the question's answer from the image using deep learning algorithms. In Visual Question Answering, most existing approaches adopt the pipeline of representing an image via pre-trained CNNs, and then using the uninterpretable CNN features in conjunction

with the question to predict the answer. Although such end-to-end models might report promising performance, they rarely provide any insight, apart from the answer, into the VQA process. In this work, we propose to break up the end-to-end VQA into two steps: explaining and reasoning. Answering textual questions from images, which is referred to as visual question answering, presents fundamental challenges to both computer vision and natural language processing communities. Significant progress has been made on VQA in recent years. The philosophy behind such a break-up is to mimic the image question answering process of human beings: first understanding the content of the image and then performing inference about the answer according to the understanding. We first generate two-level explanations for an image via pre-trained attribute detectors and image captioning model: 1). word-level: attributes, indicating individual objects and attributes the system learns from the image. 2). sentence-level: captions, representing the relationship between the objects and attributes. Then the generated explanations and question are infused to a reasoning module to predict an answer. The reasoning module is mainly composed of LSTMs.

To our knowledge, this is the first effort to break down the previous end-to-end pipeline to shed light on the VQA process. Our main contributions are summarized as follows: We propose to formulate VQA into two separate steps: explaining and reasoning. Our framework generates attributes and captions for images to shed light on why the system predicts any specific answer. We adopt several ways to measure the explanation quality and demonstrate strong correlation between explanation quality and VQA accuracy. The current system achieves comparable performance to the baselines and can naturally improve with explanation quality.

2.LITERATURE SURVEY

[1] This paper dives into the world of Visual Question Answering (VQA), where models answer questions based on images. It presents a comprehensive survey of recent advancements, analysing 25 studies published between 2020 and 2022. The focus is on the latest and most innovative works, including a brand-new classification system for VQA models based on their learning and reasoning capabilities. The authors provide an in-depth exploration of the field, highlighting the state-of-the-art techniques, common errors encountered, and promising future directions for research. Additionally, they make all reviewed studies readily available to the wider scientific community.[2] Finally, the paper evaluates the performance of the ROCK model on the KnowIT VQA dataset. The model is compared to other methods, including another video question answering system (TVQA) and human evaluators. The results show that while ROCK surpasses TVQA in accuracy, it still falls short of human performance, particularly on questions requiring external knowledge. The document further provides specific examples and analysis of the model's performance.[3] offers a fresh perspective on multimodal machine learning, focusing on the inherent 6 characteristics of different data types. It delves into the unique qualities and commonalities between various modalities like vision, audio, text, and motion. [4] the ability to answer questions that require reading and reasoning about text present within images. This capability is crucial for applications like assisting users with visual impairments, who rely on technology to understand their surroundings. [5] a task that combines natural language processing and computer vision to answer questions about images. It provides a comprehensive review of the datasets and models employed in this domain.

3.PROPOSED SYSTEM

The system that we propose to overcome the drawback of existing system is an Enhancing Human Computer Interaction which can be used for answering the question from the given image. A Convolutional Neural Network (CNN) is a type of deep neural network specifically designed for processing structured grid-like data, such as images. CNNs are particularly powerful for tasks in computer vision, such as image classification, object detection, and image segmentation. They are used to transform the input image into a fixed-size representation that can be further processed by other components of the VQA model. CNNs can learn hierarchical representations of visual data, capturing increasingly complex features as the network progresses through its layers.

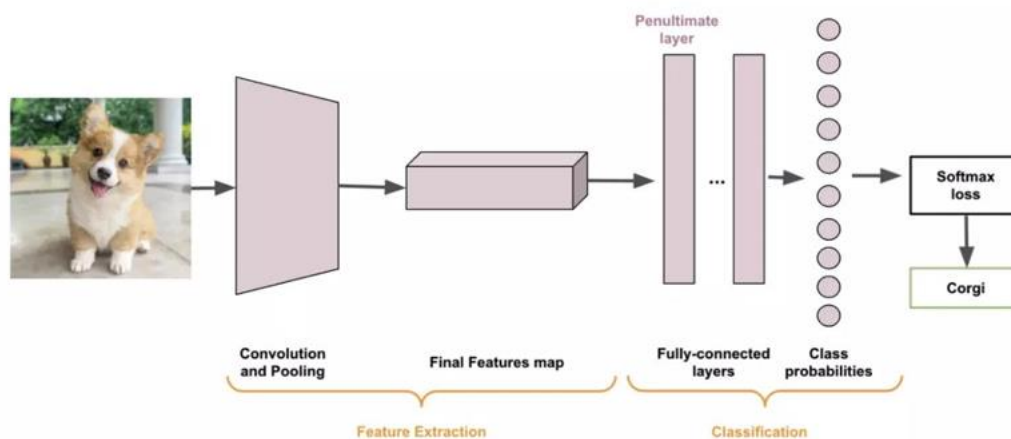


Fig1. CNN WORKING

It is the process of analyzing and understanding natural language questions in order to extract relevant information and structure from them. It involves breaking down a question into its constituent parts, such as words, phrases, and grammatical structures, and then interpreting the meaning and intent behind those components. During question parsing, various linguistic techniques and tools are often employed, including tokenization, part-of-speech tagging, named entity recognition, dependency parsing, semantic role labeling, and syntactic analysis. The ultimate goal of question parsing is to transform a raw natural language question into a structured representation that can be easily understood and processed by a computer system. In the context of Visual Question Answering (VQA), question parsing involves not only understanding the textual part of the question but also integrating visual context from accompanying images.

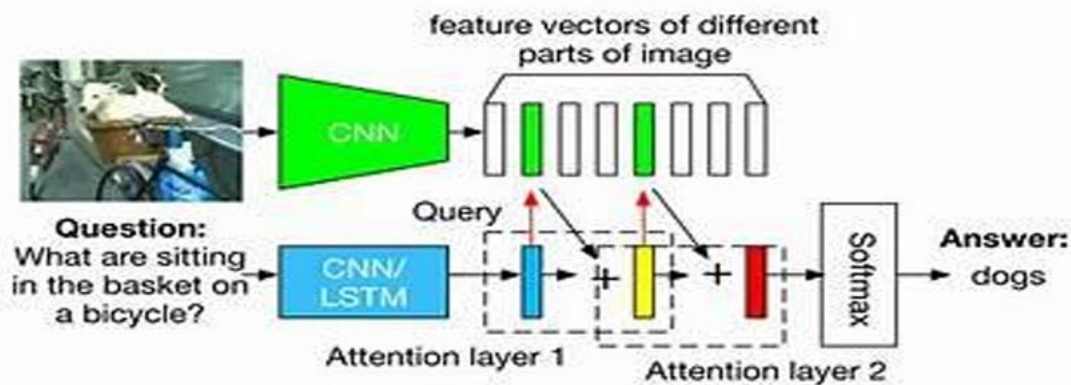


Fig2. Answer prediction

Visual Question Answering (VQA) systems typically consist of several interconnected modules, each serving a specific function in the overall process of understanding images and answering questions about them. This summary provides an overview of the key modules involved in VQA systems, including image preprocessing, question processing, multimodal fusion, and answer generation. Neural Module Network approach commences by dynamically parse the question into a canonical form (i.e. a dependency tree). Given the question under this canonical form, a layout of modules is then generated and used to construct a prediction pipeline., This pipeline will then be used to predict the answer module by module in a bottom-up manner. These modules take image features as input and produce a vector that can be combined with the encoded question to determine the answer, which is a classification problem over a built-in answer dictionary.

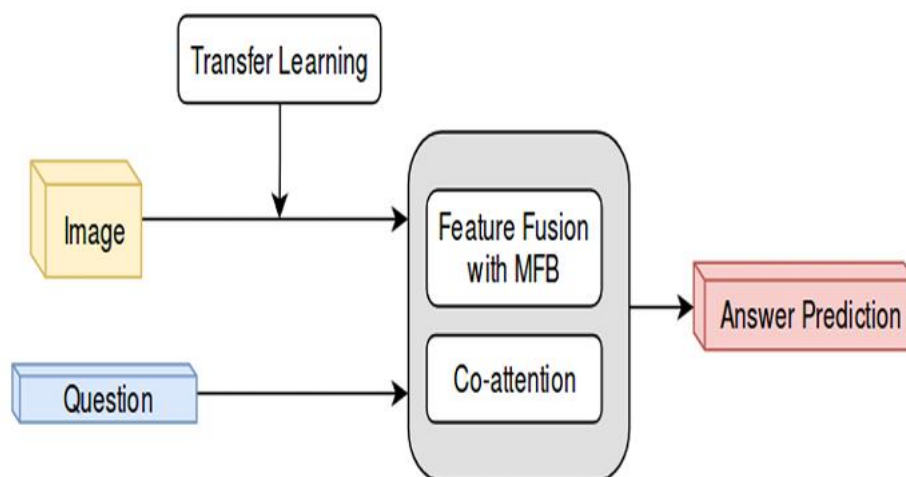


Fig3. System Architecture

4.SYSTEM RESULTS

System results are the outcome of the whole process of software testing life cycle. The results thus produced, offer an insight into the deliverables of a software project, significant in representing the status of the project to the stakeholders. The analysis of system results is an important part of system evaluation, as it provides insight into the effectiveness and efficiency of the system and helps to identify areas for improvement.

5. CONCLUSION

The proposed system Visual Question Answering represents an exciting frontier in AI, combining computer vision and natural language processing. While significant progress has been made, challenges like context understanding and bias remain. Yet, the potential for applications like image captioning and accessibility tools is vast, driving ongoing research to improve accuracy and robustness. However, several challenges persist. VQA models often struggle with complex questions requiring nuanced understanding of context, spatial relationships, and common-sense reasoning. They also exhibit biases present in training data, leading to skewed or inaccurate responses. Additionally, ensuring the robustness and interpretability of VQA systems remains an ongoing concern. Despite these challenges, the continued advancement of VQA holds great promise for numerous applications, including image captioning, content-based image retrieval, and assistive technologies for the visually impaired. Future research efforts should focus on addressing existing limitations, improving model generalization, and fostering a deeper understanding of the interplay between vision and language. Through collaborative interdisciplinary efforts, the field of VQA can continue to evolve, enabling machines to comprehend and interact with visual information in increasingly sophisticated ways.

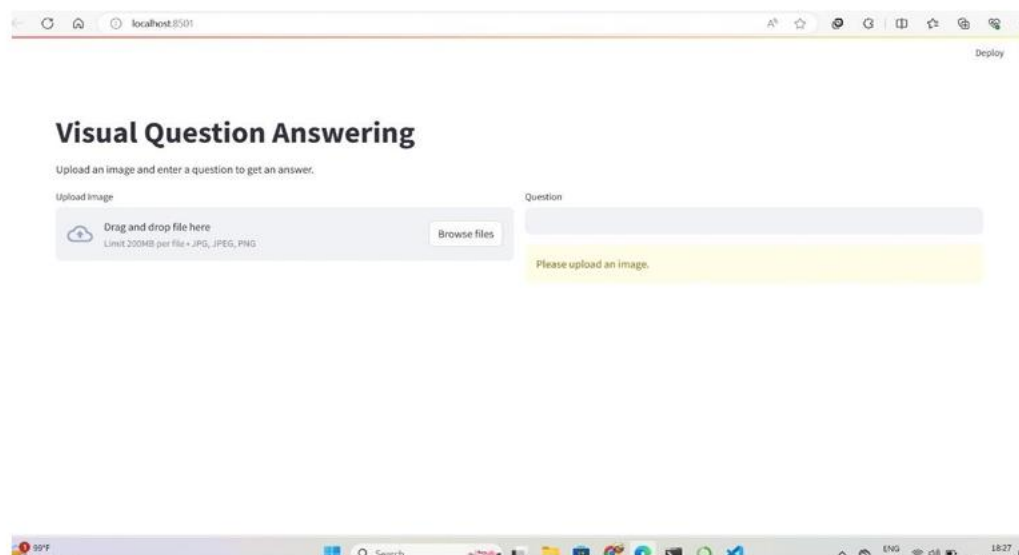


Fig4. Initial User Interface for Visual Question and Answering

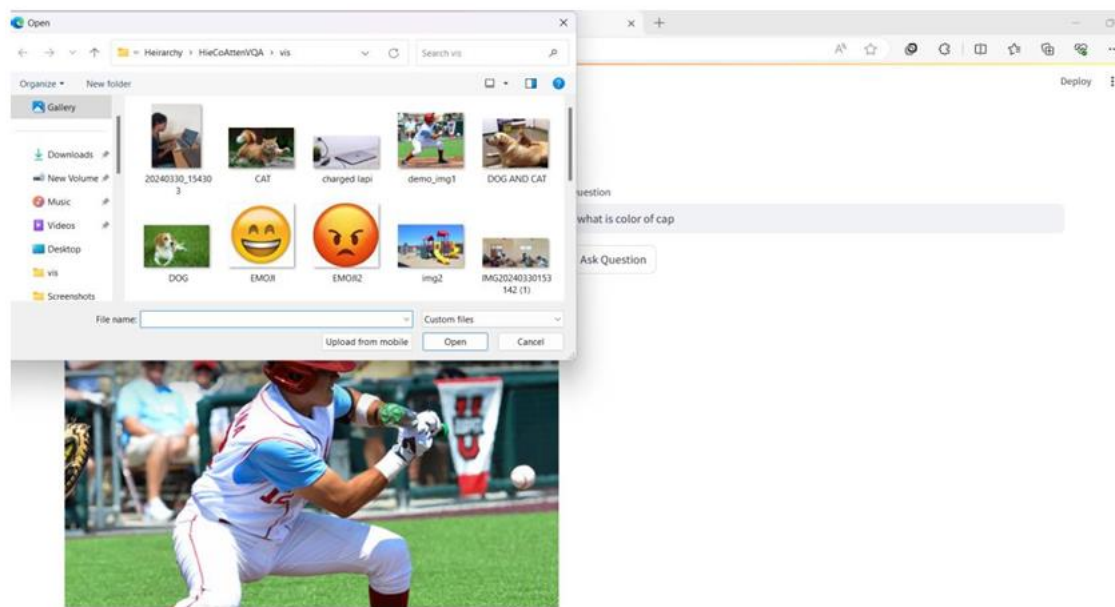


Fig5. Shows the window opened when browse files button is clicked

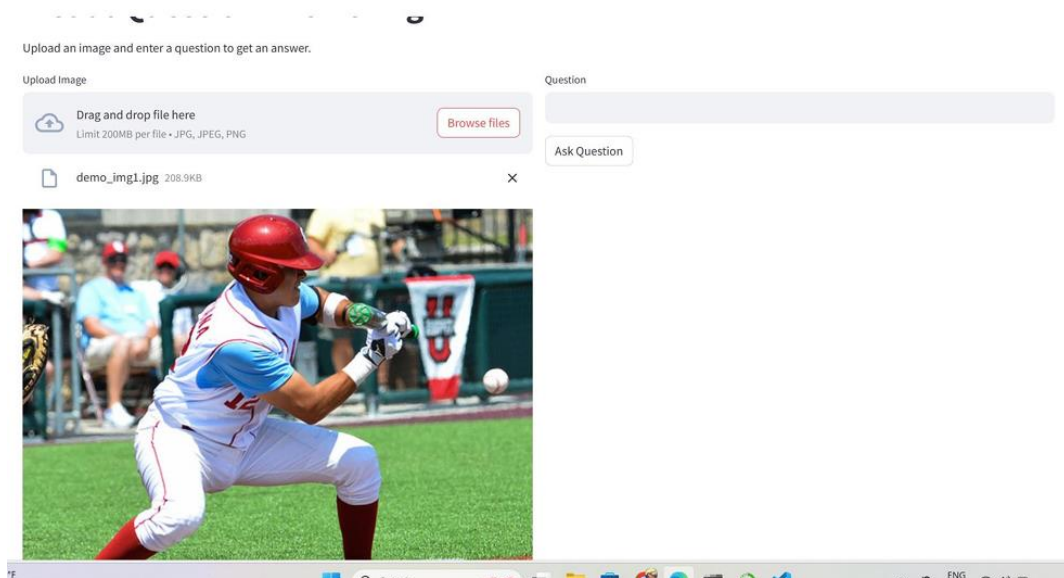


Fig6. Image is selected

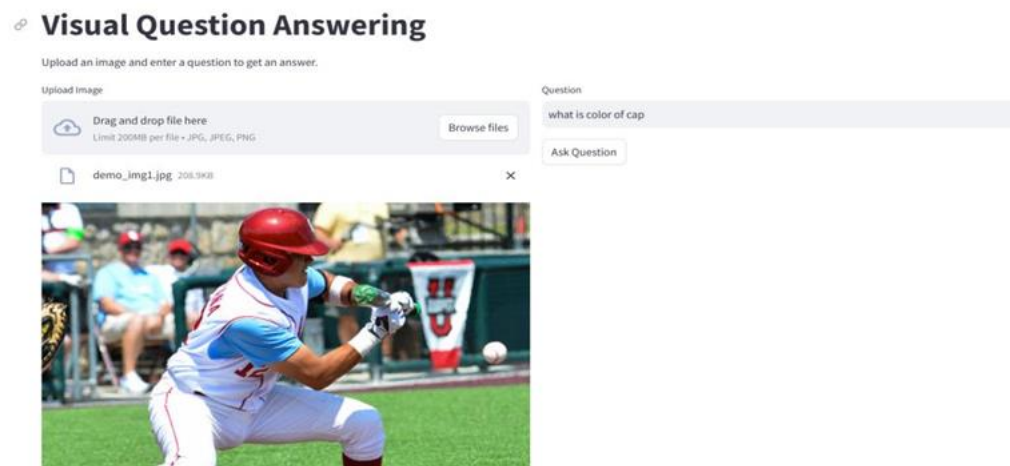


Fig7. Question is entered as colour of cap

Visual Question Answering

Upload an image and enter a question to get an answer.

Upload image

Drag and drop file here

Limit 200MB per file • JPG, JPEG, PNG

Browse files

demo_img1.jpg 208.9KB



Question

what is color of cap

Ask Question

Your Question is color of cap

Answer: red

Fig8. Colour is predicted as red

5. CONCLUSION

The proposed system Visual Question Answering represents an exciting frontier in AI, combining computer vision and natural language processing. While significant progress has been made, challenges like context understanding and bias remain. Yet, the potential for applications like image captioning and accessibility tools is vast, driving ongoing research to improve accuracy and robustness. However, several challenges persist. VQA models often struggle with complex questions requiring nuanced understanding of context, spatial relationships, and common-sense reasoning. They also exhibit biases present in training data, leading to skewed or inaccurate responses. Additionally, ensuring the robustness and interpretability of VQA systems remains an ongoing concern. Despite these challenges, the continued advancement of VQA holds great promise for numerous applications, including image captioning, content-based image retrieval, and assistive technologies for the visually impaired. Future research efforts should focus on addressing existing limitations, improving model generalization, and fostering a deeper understanding of the interplay between vision and language. Through collaborative interdisciplinary efforts, the field of VQA can continue to evolve, enabling machines to comprehend and interact with visual information in increasingly sophisticated ways.

REFERENCES

- [1] A. Agrawal, D. Batra, and D. Parikh. Analyzing the Behavior of Visual Question Answering Models. In EMNLP, 2016.
- [2] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Deep compositional question answering with neural module networks. In CVPR, 2016.
- [3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual Question Answering. In ICCV, 2015.
- [4] X. Chen and C. L. Zitnick. Mind's Eye: A Recurrent Visual Representation for Image Caption Generation. In CVPR, 2015.
- [5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In CVPR, 2009.

- [6] J. Devlin, S. Gupta, R. B. Girshick, M. Mitchell, and C. L. Zitnick. Exploring nearest neighbor approaches for image captioning. CoRR, abs/1505.04467, 2015.
- [7] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. In CVPR, 2015.
- [8] H. Fang, S. Gupta, F. N. Iandola, R. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig. From Captions to Visual Concepts and Back. In CVPR, 2015.
- [9] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. In EMNLP, 2016.
- [10] H. Gao, J. Mao, J. Zhou, Z. Huang, and A. Yuille. Are you talking to a machine? dataset and methods for multilingual image question answering. In NIPS, 2015.
- [11] Y. Goyal, A. Mohapatra, D. Parikh, and D. Batra. Towards Transparent AI Systems: Interpreting Visual Question Answering Models. In ICML Workshop on Visualization for Deep Learning, 2016.
- [12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, 2016.