Visual Question Generation Answering (VQG-VQA) using Machine Learning Models

ATUL KACHARE, MUKESH KALLA, ASHUTOSH GUPTA
Computer Science and Engineering
Sir Padampat Singhania University
Udaipur, Rajasthan
INDIA

Abstract: - Presented automated visual question-answer system generates graphics-based question-answer pairs. The system consists of the Visual Query Generation (VQG) and Visual Question Answer (VQA) modules. VQG generates questions based on visual cues, and VQA provides matching answers to the VQG modules. VQG system generates questions using LSTM and VGG19 model, training parameters, and predicting words with the highest probability for output. VQA uses VGG-19 convolutional neural network for image encoding, embedding, and multilayer perceptron for high-quality responses. The proposed system reduces the need for human annotation and thus supports the traditional education sector by significantly reducing the human intervention required to generate text queries. The system can be used in interactive interfaces to help young children learn.

Key-Words: - Visual Question Generation, Visual Question Answer, Image Feature Extraction, E-Learning System

Received: October 12, 2022. Revised: May 19, 2023. Accepted: June 11, 2023. Published: June 28, 2023.

1 Introduction

Question answering (QA) and question generation (QG) are essential tasks in communication due to progress made in various areas of machine learning. Neural networks have greatly improved the speed and accuracy of image processing tasks such as object recognition and image segmentation and natural language processing tasks such as input recognition, language generation, and question answering. The result is a multidisciplinary project known as VQA and VQG that combines these methods and combines computer vision and natural language processing techniques. VQA and VQG must make inferences between text questions and answers based on the content of related images. VQA focuses on answering questions about images, while VQG aims to generate meaningful questions based on the content of the images and the answers given. Visual Question Answering (VQA) and Visual Question Generation (VQG) are popular topics in computer vision but are often studied separately despite their intrinsically complementary relationships. This paper aims to comprehensively review visual query generation and query answering, including their methods and existing datasets.

VQG, considered complementary to VQA, has recently attracted considerable attention as a fascinating problem. Its objective is to generate meaningful questions based on input images. This task involves image comprehension and natural language generation, often employing deep learning techniques. In VQG, the first step is comprehending the picture and generating a coherent sequence of texts that constitute syntactically and semantically valid questions. Image comprehension involves successfully detecting objects, classifying objects, labeling them, identifying relationships among objects, understanding the scene, and classifying the scene.

Visual question-answering systems aim to respond to natural language questions based on visual input accurately. A broader perspective of this problem is to develop systems that can comprehend image content in a human-like manner and effectively communicate about it using natural language. This task is challenging as it requires the interaction and synergy between image-based and natural language models. It is widely regarded as a crucial milestone in the development of artificial intelligence and represents the effort to make computers as intelligent as humans. Some researchers have even proposed using visual question answering as a benchmark for evaluating AI systems' capabilities, like the Turing Test concept, [1].

To provide an overview of the subproblems involved in visual question answering, consider the following examples in Table 1:

Solving these challenges involves four main steps: Image Featurization, Question Featurization, Joint Feature Representation, and Answer Generation. The remaining paper is organized as follows: section 2 briefly describes the literature review on VQG and VQA. The proposed automated visual question-answer system is presented in section 3. The dataset, data preparation, exploratory data analysis, and results are reported in section 4. Finally, section 5 concludes the work.

Table 1: Computer Vision Task required to be solved by VQA, [2]

CV Task	**Representative VQA Question**
Object Recognition	What is in the image?
Object Detection	Are there any books in the image?
Attribute Classification	What color is the book cover?
Scene Classification	Is it a day or night?
Counting	How many chairs are there?
Activity Recognition	What the person is doing?
Spatial Relationship among Objects	What is on the desk between the bottle and the computer?

2 Review of Literature

A neural design for answering natural language questions on the illustrations using an amalgamation of CNN and LSTM is reported, [3]. They performed experiments and concluded that the system with Images performed better than the previous system without images and also proposed two metrics Average Accord which considers the human divergence, and Min Accord, which captures the dispute in Human QA.

A free-form and open-ended VQA that can answer any conversational language inquiry about a picture by combining VGGNet alongside LSTM for image and question processing is created, [4]. It can answer ”wh” questions along with the knowledge representation questions. The paper suggested a need for a task-specific dataset to answer the practical VQA application questions more efficiently.

In, [5], the authors fabricated the visually trained queries with distinct varieties for an individual picture using VGGNet and LSTM. Authors experiment with VisualQA and Visual7W datasets, generating captions and questions for comparing existing systems. They found that the system could generate a more significant number of questions than the previously existing systems. Even the questions generated were reasonable and grammatically well-formed.

A unique neural network architecture that improves image-based query generation using a dependency analysis tree and a CNN, [6]. The proposed model, a parse-tree-directed residual network (PTGRN), comprises three collective modules. The attentiveness module uses a native visible exhibit, the closed residual composition module aggregates previously minified evidence, and the parse-tree-driven dissemination component forwards the decreased suggestion alongside the parse tree.

In, [7], a system for generating free-form unconstrained video questions and answers using an attention model with bidirectional LSTM is developed. The researchers created a video quality control dataset using automatic query generation and proposed two alerting mechanisms. Sequential Video Attention computes video attention while preserving the sequential structure of the question, and Temporal Question Attention collects question attention for each video frame.

A new approach to integrate advanced concepts in a CNN-RNN approach. The system includes image analysis that learns to combine image and semantic properties using CNN successfully proposed in, [8]. The linguistic segment uses LSTM to learn the association between the attribute vector and the word sequence. The authors use ImageNet’s pre-trained VGGNet for initialization, adapting for multi-label datasets and attribute predictions. They use semantic attribute predictive values instead of natural features. A partitioned CNN generates proposal regions, aggregating output for high-level image depiction.

In, [9], authors uses a dual learning framework for joint learning of VQG and VQA tasks. The system uses GRNN and NeuralTalk2 model. We use two agents using dual-learning-adjusted pre-trained models. The proposed model regularly outperforms existing VQA techniques, which suggests that dual learning offers a natural closed-loop strategy for both VQG and VQA tasks and that the VQA task supports the development of the VQG job’s performance.

In, [10], the authors proposed introducing two tasks of VQA improvisation for VQG quality assurance using the RNN and LSTM model. They formulated VQG and VQA as reverse processes, separating the system and the duality regulator. The proposed method reconstructs the VQA model into its dual VQG form, allowing us to train a single model with two conjugate tasks. The experimental results show that iQAN, with two training courses, learned the interaction between answers, questions, and images bi-directionally.

A fine-grained picture and query architecture that allows deep neural networks and a co-attention framework to identify interests devised in, [11]. They proposed that to support VQA, we must solve a trio of issues: distinctive fine-grained illustrations across the picture and the query; multidisciplinary feature amalgamation which can record the intricate relationships of multidisciplinary traits; and inherent response forecasting that is capable of taking note of complicated associations across various distinct responses to an identical concern. It is feasible to successfully eliminate irrelevant characteristics and obtain distinctive characteristics for photos and queries employing an

integrated awareness paradigm.

In, [12], authors developed a system that automatically generates visual question-answer pairs by generating and answering an image given as input. In this work, VQG and VQA are performed sequentially. The VQG model uses CNN and RNN, while the VQA model uses CNN to extract visual features, using the question embedding and visual features as input to obtain the answer. It eliminates the need for human intervention in data input. This method creates an entity pair and an attribute based on the question and the answer. However, the system's limitation is that it can only give a one-word answer to a question.

In, [13], authors investigated the association between top-down visibility and text production to determine whether accurate visibility responses benefit text generation. The application consists of text-based top-down visibility and visibility-based VQG mesh. The first component takes an image and a query as input and creates a visibility map correlating with the query. The second component takes as input an image and visibility, which generates an appropriate query based on the most critical areas. A two-supervised network with dynamic parameter prediction was proposed. This method uses the probabilistic relationship between top-down detection and text production while dynamically predicting parameters to fully encode the given text into a convolutional networking parameter to encode the given text into a convolutional network fully. The proposed top-down explanatory method correlates well with human attention.

The impact of prediction parameter modification on VQA was examined in, [14], using Stochastic Gradient Descent (SGD) with pre-trained faster R-CNN and LSTM models. Benchmark VQA datasets used for evaluation. The proposed method effectively handles various inquiries requiring different levels of semantic understanding beyond simple image content questions. The VQA model achieves improved accuracy and generates answers for free-form unrestricted questions based on images. It consists of four components: Image Features, Question Features, Parameter Prediction, and training using the Stochastic Gradient Descent Approach.

In, [15], authors discussed the different textual and visual image feature extraction methods and the single and multi-hop attention models for Visual Question answering systems. In visual feature extraction, authors have discussed from LeNet model with 2 CNN layers for handwritten digit recognition to complex VGG, GoogleNet, and ResNet models. In textual feature extraction, we have a family of RNNs, including LSTM, GRU, and Bi-LSTM, in the case of fusion strategies broadly classified as vector operators, neural networks, or bilinear pooling. Also, the authors have highlighted the single and multi-hop attention models for textual and visual channels such as LSTM-Attention and MLAN.

In, [16], the authors built a model that increases the collective knowledge between images, presumed responses, and produced questions. They introduced a continuous latent space that varied in the expected response project to deal with differences in individual natural language codes. They normalize this concealed space with a subsequently suppressed space that guarantees similar response groups. In addition, if the system does not know the expected answer, the second hidden space can contain objects and materials. It can generate targeted questions to elicit, which quantifies the model's ability to store information about expected response categories, resulting in more diverse, targeted questions.

A visual question-answering system for Remote Sensing Data with the help of the Convolution and Recurrent Neural Networks with a bitwise merger of both designed in, [17]. They built two datasets using low- & high-resolution images of images/question/answer triplets. CNN and RNN combine for visual and natural language question prediction, fused by point multiplication and OpenStreetMap for QA generation. Nevertheless, OSM's limitations caused the system's accuracy degradation. Also, the answers were limited compared to the traditional VQA dataset.

In, [18], authors presented a current standardized data source containing queries prepared by human annotators while considering inquiries people would ask multimedia virtual assistants. A pre-trained CNN and Text encoder LSTM model analyze image content and metadata to generate meaningful queries. They presented a new dataset that included nearly four times the number of queries as the OK-VQA dataset. In addition, their approach was tested against industry standard evaluation measures such as BLEU, METEOR, ROUGE, and CIDEr to determine the relevance of the questions created with the questions submitted by users. They also examined the diversity of produced queries using generative strength and originality criteria and discovered that they performed better.

Complex tasks for classification and response generation were transformed into several simple tasks by, [19]. They used pre-trained ResNet152 for image extraction and three types of attachments (location, segment, and character) for text processing. The authors also used a multi-terminal self-controlled transformer to reduce the computational cost. On the ImageClef2029 VQA-Med dataset, the suggested system, CGMVQA, was tested. It demonstrated higher accuracy for fundamental questions, making it appropriate for early medical students and patient care.

In, [20], authors introduced an innovative

response-oriented approach called Radial Graph Convolutional Network (Radial-GCNy. The paper proposed a method that identifies a set of candidate regions in an image, then identifies a core response region within those regions. Construct a radial graph, and use graph convolution for contextualizing graphic and semantic descriptions. The mean association function is then fed to a universal LSTM interpreter to generate a profound query correlated with the vision and the response. The central concept of the proposed system is that no extensive image analysis is required to prepare inputs for query generation. The model effectively uses graphical and semantic knowledge to determine the location of the response area, which complicates space and time.

In, [21], authors improved the fairness of the answers in terms of ethically sensitive attributes with the help of faster R-CNN and Gated Recurrent Network. The system consists of two basic models: the SAP and VQA modules. VQA module primarily generates all possible (fair & unfair) answers, whereas the SAP module predicts the sensitive attributes of answers. Answers of both modules are combined using a debiased fusion scheme to yield the final solution. The system's limitation is that it was studied only on gender attributes. Hence the generalizability of the system is not evaluated.

The VQA model presented by, [22], is based on detecting visual relationships between multiple objects. Word vector similarity concepts are used in layout models, swapping original object features for image attributes and representing aspects and relational predicates. The classifier merged the components of the image feature and the question vector as input for proving the answer. The system comprised an object-detection model and an object-related estimation model. An aspect ratio model was applied to increase the model's generalizability and aid in the conclusion of linkages.

An intelligent manufacturing system using VQA using the ImageNet dataset and ResNet pre-trained for LSTM propsed, [23]. The conceptual model consists of five parts: the physical HMC system, the virtual HMC system, the service system of the HMC system, the DT data of the HMC system, and the link between the four deployed components. The proposed VQA model, a video-text association network, can understand visual and textual information. It can answer simple multiple-choice questions and create a sentence to answer open-ended questions. In this way, people and machines can work together more conveniently and efficiently.

In, [24], the authors proposed an attention-based mechanism for generating visual queries using a simple RNN and LSTM encoder-decoder model with a DNN-based attention mechanism. The paper compares the result of a simple encoder-decoder model with an attention-based model. The proposed model is efficient and effortless. The downside is that the system only focuses on valid questions about color specifications.

A difficulty-driven generative network was proposed by, [25], where an automatic question generator generates questions with difficulty levels adjusted according to the user's skill and experience using RCNN and LSTM. They used the training domain difficulty index to identify a difficulty variable demonstrating the intricacy intensity of the questions and combined it with our model to power the generation of questions with manageable difficulties. The difficulty management mechanism combines the difficulty information with the decipherer commencement, and each time step contributes to managing the intricacy level of the created difficulties.

In, [26], the authors purported the knowledge-based Visual Question Generation model. They have used pre-trained models to generate the object-level features of objects in images. Then Encoder model combines visual object level and non-visual knowledge information. The overall system mainly consists of 4 different components: the visual concept feature extractor, knowledge feature extractor, target object extractor, and decoder module. In Visual extraction, not only are image features extracted, but with the help of a Graph Neural Network, the spatial relationship between multiple objects is detected and represented using a Sparse Graph. The Answer-Aware module, a part of the Knowledge feature, is vital in finding non-visual information.

A solution involving a transformer-grounded sight and verbal model proposed by, [27]. The researchers used the Swin Transformer encoder to produce a multiscale visual representation. This representation serves as a prefix that helps Generative Pretrained Transformer-2 decoders generate several queries in paragraph form, effectively parsing the rich visual information in Remote sensing scene captures. An audio decoder was optimized using the RS dataset for related queries from photos. The model was accessed using two VQA data sources, and a new, fully human-annotated TextRS-VQA data source was introduced to improve the assessment of VQG models.

In, [28], authors used a fully automated method to create the first comprehensive His VAQA (Arabic Visual Question Answering) dataset. This dataset consists of approximately 138,000 triplets of photo-question-answer (IQA) pairs focused on yes/no questions related to real-world photos. They created their database structure and his IQA ground truth-generating technique exclusively for the automated compilation of VAQA datasets. The five components of the system are answer prediction, question prepa-

ration, visual feature extraction from text, and feature fusion and answer prediction. The authors identified the most efficient strategy for Arabic queries during the question preprocessing and presenting stages of this study, as it was the first study to investigate VQA in Arabian. To do this, they created 24 Arabic VQA models that tested four His LSTM networks using various question tokenization schemes, 3-word embedding techniques, and architectural designs. To assess the efficacy of several Arabic VQA models authors did a thorough performance analysis of the VAQA dataset. According to the trial results, the Arabic VQA model performed between 80.8% and 84.9%.

VQG and VQA can be combined to form a dual system that eliminates the need for human annotators and avoids relying on image captioning. The VQG component generates questions based on input images. It learns to generate questions that are relevant to the visual content of the images. By using fine-grained parameters, the VQG system can be trained to produce more robust and detailed questions, capturing various aspects of the visual information. The VQA component answers the questions generated by the VQG system. It takes the input image and the corresponding question and produces an answer. By using fine-grained parameters, the VQA system can be trained to provide more accurate and detailed answers, considering subtle visual cues and nuances. The combined VQG and VQA system can be trained using large datasets that contain paired images, questions, and answers. This eliminates the need for manual annotation, as the questions and answers can be automatically generated and paired with the images. The system can be trained end-to-end, optimizing both the question generation and answer prediction tasks simultaneously. By using fine-grained parameters, the system can capture more nuanced and detailed information from the images, leading to more robust performance. Fine-grained parameters allow the system to focus on specific visual attributes, objects, or relationships, improving its ability to understand and generate relevant questions and accurate answers. Overall, the dual system of VQG and VQA, combined with fine-grained parameters, provides a self-contained framework for generating questions and answering them based on visual input. It reduces the reliance on human annotators and avoids the limitations of image captioning, while also enabling the system to achieve higher robustness and accuracy in understanding and processing visual information.

3 Proposed System

The proposed system will combine the VQG and VQA systems to first create the questions from the input images provided and then create the answers for those questions, which will be used later in the system. In the Visual Question Generation (VQG) system shown in Figure 1, LSTM produces questions, and the pre-trained CNN model extracts image features—the COCO and VQA dataset used for training and evaluation.

The LSTM method uses parameters to train, with each image and query as a record in the embedding. The method repeats the process to reveal previous embeddings and predict future output states. The highest significant probability predicts the word and the generated embedding matches the word with the highest probability in each successive layer. The generated questions are then output using the generated words.

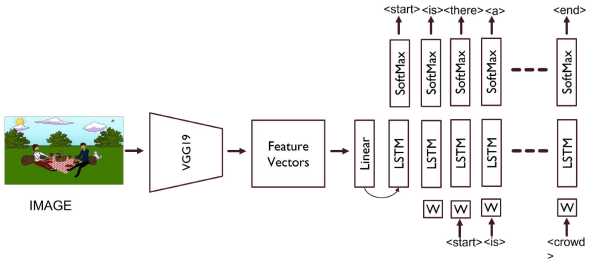

Figure 1: Visual Question Generation Model

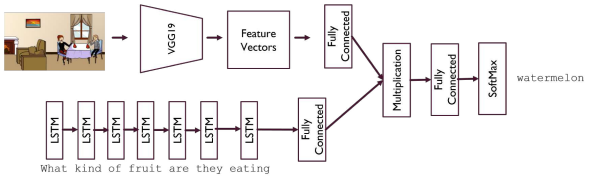

Figure 2: Visual Question Answering Model

Figure 2 indicates the high-level baseline architecture of our own VQA system. The image has a 224x224 scale. A convolutional neural network (CNN) made up of VGG-19 receives the scaled image as input.

CNN outputs a characteristic vector that encodes the content of the photo, known as picture embedding. The query is passed to the embedding layer, creating compact, whole-dimensional embedding vectors. So, they are first projected to an equal range of dimensions using the corresponding connected layers (linear transformations) and then blended with pointwise multiplication (multiplying the values within the corresponding dimensions). The very last degree of the VQA model is a multilayer perceptron with a final SoftMax nonlinearity that outputs the rating distribution for each of the highest quality ok (1000) responses. Changing the responses to a k-manner category assignment allows us to educate a VQA model

on using the go-entropy loss among the generated reaction distribution and the floor reality.

4 Experimental Setup

4.1 Dataset

We have used the VisualQA (VQA) and COCO datasets to evaluate the proposed system.

VQA Dataset: The dataset contains the 20000 images in the training set and 60000 question-answers pair for training images. Also, we have 10000 validation images and 20000 test images. The dataset contains 30000 question-answers pair for validation.

COCO Dataset: The dataset contains the 82783 images in the training set and 4437570 question-answers pair for training images. Also, we have 40504 validation images and 81434 test images. The dataset contains 2143540 question-answers pair for validation.

4.2 Data Preparation

Since unstructured and structured data are used, it is necessary to properly administer generator prototype management in multimodal systems for generating solutions.

4.3 Exploratory Data Analysis

The Figure 3, Figure 4, Figure 5, and Figure 6 show some exploratory data analysis for the dataset.

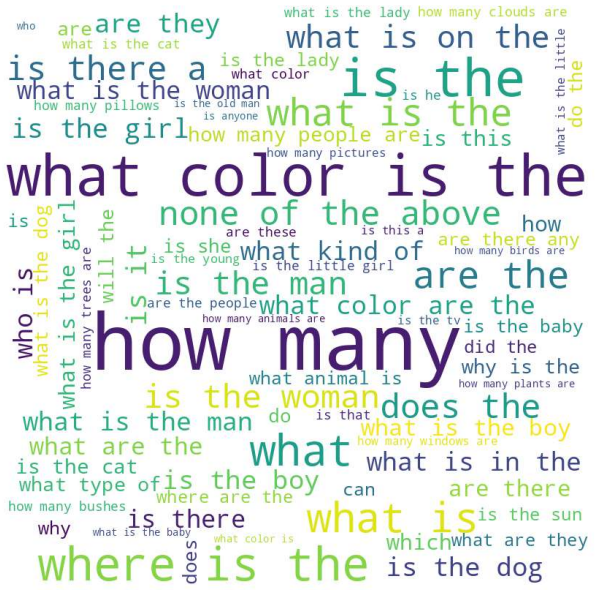

Figure 3: Word Cloud on Question-Type

4.4 Result

The Table 2 below shows the system's accuracy with different batch sizes and datasets.

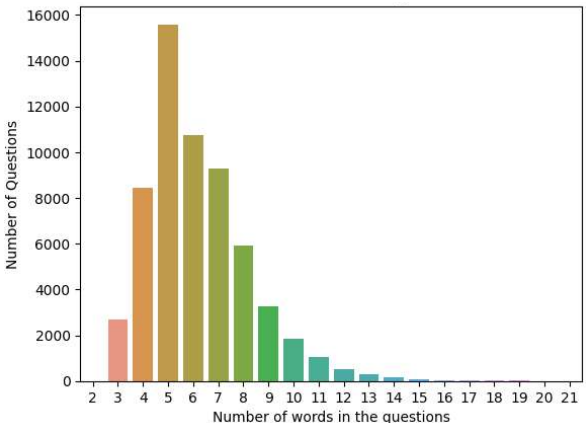

Figure 4: Length of Questions

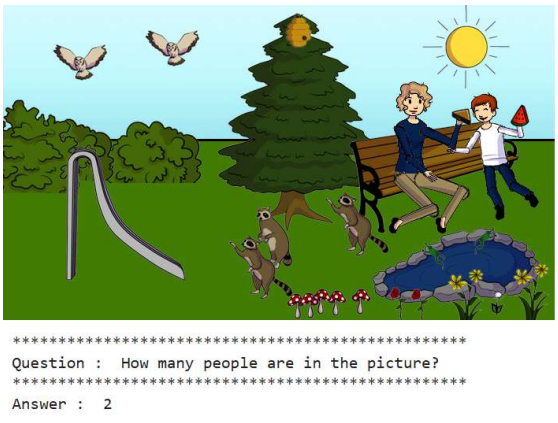

Figure 5: Sample Triplet (Image, Question, Answer)

Figure 6: Output of the System

Table 2: Accuracy with different modes, datasets, and batch sizes

Model	Batch Size	Accuracy		
		Top 1	Top 3	Top 5
VQA with VGG19	16	0.5108	0.7904	0.8456
	32	0.5165	0.7917	0.8459
	64	0.5224	0.7943	0.8514
	128	0.5136	0.7914	0.8443
	256	0.5251	0.7966	0.8521
	512	0.5151	0.7928	0.8473
VQA with ResNet 152	256	0.4883	0.7598	0.8206
	512	0.2520	0.4856	0.5447
COCO with VGG19	64	0.3509	0.6566	0.7179
	128	0.3566	0.6550	0.7157
	256	0.3603	0.6608	0.7220
	512	0.3701	0.6663	0.7261

5 Conclusion

An interdisciplinary field called VQA combines verbal expertise with visual information. The importance of this field rests in how it combines language comprehension with visual interpretation. However, a computerized tool combining VQG and VQA capabilities is not available. In our project, we developed a system that responds to questions that our VQG model creates using our VQA model. According to our encouraging results, with enough training data, the system should be able to produce questions and endorse them by employing a powerful question-answering component.

Our long-term goals include improving the system's capabilities and addressing its flaws. We aim to make the VQA technique flexible to produce whole sentences because the existing VQA tool only allows users to respond to questions generated by VQG with a single word. By incorporating emotion recognition, motion detection, and event comprehension into photos and creating pertinent answers based on these features, we hope to improve the system. In addition, we want to improve the precision of the existing VQG, and VQA approaches to create more naturally occurring query-response pairings.

References:

[1] D. Geman, S. Geman, N. Hallonquist, and L. Younes, "Visual turing test for computer vision systems," *Proceedings of the National Academy of Sciences*, vol. 112, no. 12, pp. 3618–3623, 2015.

[2] S. Manmadhan and B. C. Kovoor, "Visual question answering: a state-of-the-art review," *Artificial Intelligence Review*, vol. 53, pp. 5705–5745, 2020.

[3] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A neural-based approach to answering questions about images," in *Proceedings of the IEEE international conference on computer vision*, pp. 1–9, 2015.

[4] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

[5] S. Zhang, L. Qu, S. You, Z. Yang, and J. Zhang, "Automatic generation of grounded visual questions," *arXiv preprint arXiv:1612.06530*, 2016.

[6] Q. Cao, X. Liang, B. Li, and L. Lin, "Interpretable visual question answering by reasoning on dependency trees," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 3, pp. 887–901, 2019.

[7] H. Xue, Z. Zhao, and D. Cai, "Unifying the video and question attentions for open-ended video question answering," *IEEE Transactions on Image Processing*, vol. 26, no. 12, pp. 5656–5666, 2017.

[8] Q. Wu, C. Shen, P. Wang, A. Dick, and A. Van Den Hengel, "Image captioning and visual question answering based on attributes and external knowledge," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1367–1381, 2017.

[9] X. Xu, J. Song, H. Lu, L. He, Y. Yang, and F. Shen, "Dual learning for visual question generation," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, IEEE, 2018.

[10] Y. Li, N. Duan, B. Zhou, X. Chu, W. Ouyang, X. Wang, and M. Zhou, "Visual question generation as dual task of visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6116–6124, 2018.

[11] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, "Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 12, pp. 5947–5959, 2018.

[12] S. Nahar, S. Naik, N. Shah, S. Shah, and L. Kurup, "Automated question generation and answer verification using visual data," *Modern Approaches in Machine Learning and Cognitive*

Science: A Walkthrough: Latest Trends in AI, pp. 99–114, 2020.

[13] S. He, C. Han, G. Han, and J. Qin, "Exploring duality in visual question-driven top-down saliency," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 7, pp. 2672–2679, 2019.

[14] S. Jha, A. Dey, R. Kumar, and V. Kumar, "A novel approach on visual question answering by parameter prediction using faster region based convolutional neural network," *IJIMAI*, vol. 5, no. 5, pp. 30–37, 2019.

[15] D. Zhang, R. Cao, and S. Wu, "Information fusion in visual question answering: A survey," *Information Fusion*, vol. 52, pp. 268–280, 2019.

[16] R. Krishna, M. Bernstein, and L. Fei-Fei, "Information maximizing visual question generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2008–2018, 2019.

[17] S. Lobry, D. Marcos, J. Murray, and D. Tuia, "Rsvqa: Visual question answering for remote sensing data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 12, pp. 8555–8566, 2020.

[18] A. Patel, A. Bindal, H. Kotek, C. Klein, and J. Williams, "Generating natural questions from images for multimodal assistants," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2270–2274, IEEE, 2021.

[19] F. Ren and Y. Zhou, "Cgmvqa: A new classification and generative model for medical visual question answering," *IEEE Access*, vol. 8, pp. 50626–50636, 2020.

[20] X. Xu, T. Wang, Y. Yang, A. Hanjalic, and H. T. Shen, "Radial graph convolutional network for visual question generation," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 4, pp. 1654–1667, 2020.

[21] S. Park, S. Hwang, J. Hong, and H. Byun, "Fairvqa: Fairness-aware visual question answering through sensitive attribute prediction," *IEEE Access*, vol. 8, pp. 215091–215099, 2020.

[22] Y. Xi, Y. Zhang, S. Ding, and S. Wan, "Visual question answering model based on visual relationship detection," *Signal Processing: Image Communication*, vol. 80, p. 115648, 2020.

[23] T. Wang, J. Li, Z. Kong, X. Liu, H. Snoussi, and H. Lv, "Digital twin improved via visual question answering for vision-language interactive mode in human–machine collaboration," *Journal of Manufacturing Systems*, vol. 58, pp. 261–269, 2021.

[24] C. Patil and A. Kulkarni, "Attention-based visual question generation," in *2021 International Conference on Emerging Smart Computing and Informatics (ESCI)*, pp. 82–86, IEEE, 2021.

[25] F. Chen, J. Xie, Y. Cai, T. Wang, and Q. Li, "Difficulty-controllable visual question generation," in *Web and Big Data: 5th International Joint Conference, APWeb-WAIM 2021, Guangzhou, China, August 23–25, 2021, Proceedings, Part I 5*, pp. 332–347, Springer, 2021.

[26] J. Xie, W. Fang, Y. Cai, Q. Huang, and Q. Li, "Knowledge-based visual question generation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 11, pp. 7547–7558, 2022.

[27] L. Bashmal, Y. Bazi, F. Melgani, R. Ricci, M. M. Al Rahhal, and M. Zuair, "Visual question generation from remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 3279–3293, 2023.

[28] S. M. kamel, S. I. Hassan, and L. Elrefaei, "Vaqa: Visual arabic question answering," *Arabian Journal for Science and Engineering*, pp. 1–21, 2023.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself
No funding was received for conducting this study.

Conflicts of Interest
The authors have no conflicts of interest to declare that are relevant to the content of this article.