Proceedings of the International Conference on Sustainable Computing and Data Communication Systems (ICSCDS-2023)
IEEE Xplore Part Number: CFP23AZ5-ART; ISBN: 978-1-6654-9199-0

2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS) | 978-1-6654-9199-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICSCDS56580.2023.10104870

Visual Questions Answering Developments, Applications, Datasets and Opportunities: A State-of-the-Art Survey

Harsimran Jit Singh
University of Petroleum & Energy Studies
Dehradun, India
harry20000412@gmail.com

Gourav Bathla
University of Petroleum & Energy Studies
Dehradun, India
gouravbathla@gmail.com

Munish Mehta
Lovely Professional University
Phagwara, Punjab, India
munishmehta1@rediffmail.com

Gunjan Chhabra
Department of Computer Science and Engineering,
Graphic Era Hill University, Dehradun, India
chhgunjan@gmail.com

Pardeep Singh
Department of Computer Science
and Engineering, Graphic Era Hill
University, Dehradun, India
pardeep.maan@gmail.com

Abstract

Visual Question Answering (VQA) is an emerging field in Artificial Intelligence (AI) that aims to enable machines to understand and answer questions about visual content. In this survey paper, current state-of-the art research is extensively surveyed to highlight limitations and futuristic opportunities. Visual QA systems use Natural Language Processing and machine learning techniques to understand and respond to questions posed by users. The paper then reviews the recent advances in neural network-based models and pre-trained language models. The paper also discusses the challenges facing visual QA systems, including the need for large-scale training data, the ability to handle complex and open-ended questions, and the need for robust evaluation metrics. Further, different types of datasets and evaluation metrics used in the literature are summarized, as well as the challenges and open research problems that remain to be addressed. Overall, it is concluded that VQA is a challenging task that requires a combination of visual understanding and natural language processing skills, and that there is still much scope for improvement in terms of accuracy and generalization.

Keywords— Visual Question Answering, Machine Learning, Deep Learning, Natural Language Processing

I. INTRODUCTION

In recent years, significant improvements have been made in a number of machine learning applications. Neural networks are being used more frequently for Computer Vision tasks like object detection and image segmentation as well as Natural Language Processing (NLP) activities including entity identification, language synthesis, and question answering in order to do them more rapidly and reliably. The AI community has recently become interested in the method of answering visual queries. It is a difficult task to recognize items in an image and then map a textual query for that object in the image if a user asks a question about the content that is available in that image. For instance, if any user ask that how may red colored cars are present in image. As query is in text and query is asked for image, this is quite complex to count objects in image.

A vision - based system that really can respond to questions about images in plain language was thought to be unachievable until recently. However, there has been a considerable improvement in developing systems with these characteristics since 2014. A system is given a text-based question about an image in the vision - based task known as Visual Question Answering (VQA), and it is expected to infer the answer. A large number of sub-problems are covered by questions in computer vision, such as:

(i) Identifying: the objects in an image.
(ii) Detecting objects – Is there a car in the picture?
(iii) Classification of attributes – What color is the car?
(iv) Scene categorization: Is it raining?
(v) Counting: How many cars are seen in the picture?

Beyond this, even trickier questions could be posed, including those concerning the spatial relationships between different items, like "what is the connection between the table and the sofa?" as well as rational questions. A powerful VQA system must be able to analyse images and handle a number of common computer vision tasks. Numerous applications for VQA are feasible.

VQA is a young discipline that seeks to automatically reply to open-ended inquiries about a particular image. VQA serves as the primary means of communication between the NLP and Computer Vision fields, along with image captioning. Other sub-tasks at this junction include question reasoning, object identification, object recognition, etc. VQA requires a complete knowledge of the images and questions as a result. The inability of visual question-answering techniques to respond to inquiries that need information that is not included in the image is one of its limitations. These inquiries are perhaps the most fascinating because people frequently inquire about topics that are challenging to infer from a picture. CNN are used for image recognition, while RNN are used for NLP, before the results are combined to generate the final response, as illustrated in Figure 1. Deep learning is another important field of research. This is so because techniques integrating picture identification and natural language processing are needed for visual question answering.

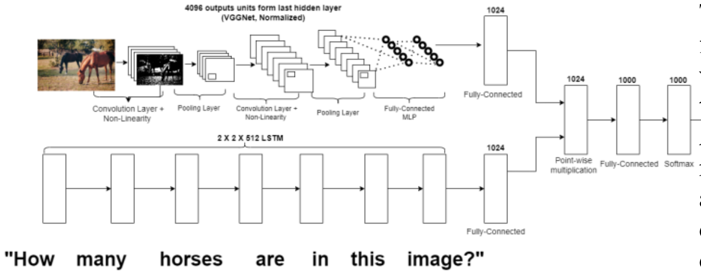
4096 outputs units form last hidden layer
(VGGNet, Normalized)
1024
Convolution Layer + Non-Linearity
Pooling Layer
Convolution Layer + Non-Linearity
Pooling Layer
Fully-Connected MLP
Fully-Connected
1024
1000
1000
"2"
Point-wise multiplication
Fully-Connected
Softmax
2 X 2 X 512 LSTM
1024
Fully-Connected
"How many horses are in this image?"

Fig. 1. Combining RNN and CNN for VQA

The contributions of this paper are as follows.

(i) Comprehensive review of state-of-the-art research works in VQA
(ii) Comparative analysis of existing research works
(iii) Review of Recent Neural networks and NLP based techniques for VQA
(iv) Applications of VQA

The remaining of the paper is organized as follows. In Section 2, traditional techniques of VQA are presented along with advantages and limitations of these research works. Section 3 covers recent developments in VQA. The significant applications of VQA are discussed in Section 4. Finally, Section 5 concludes the paper.

II. TRADITIONAL TECHNIQUES FOR VISUAL QUESTION ANSWERING

The issue in VQA is to generate a natural language response to a query about an image, using a large dataset of real-world images and various types of questions, including yes/no questions, numerical questions, and other queries. One example of a VQA dataset is the Hasan, SA dataset, which was first introduced in 2016 and is updated annually. Traditional techniques for VQA involve using a combination of computer vision and NLP methods. One common approach is to use Convolutional Neural Network (CNN) to extract features from the image, and then use Recurrent Neural Network (RNNs) to process the question and generate a response. Another approach is to use a joint embedding space to represent both the image and the question, and then use a similarity measure to find the most likely answer. These methods can be trained using supervised learning on large datasets of images and corresponding questions and answers. The task was accomplished using a variety of techniques, but the majority of them relied on deep learning techniques that combined the use of word embedding with various Recurrent Neural Networks (RNN) for text embedding and feature extraction and CNN for feature extraction from visual data, along with more complex techniques like attention mechanisms. (VQA) is a developing field, aims to automatically respond to open-ended questions on a specific image [1]. Along with image captioning, VQA acts as the main link between the Natural Language Processing (NLP) and Computer Vision areas. This intersection includes a number of other sub-tasks, such as question reasoning, object identification, object recognition, etc. Because of this, VQA demands a thorough understanding of the visuals and questions. visual question-answering techniques' incapacity to respond to inquiries requiring details not present in the image (such as Who is the man in the picture? How are they addressed? The recovered visual and verbal feature vectors are frequently merged and included via concatenation, element-wise summation, or product in order to infer a response. On the basis of the salient components in the photographs, Anderson et al. developed a bottom-up [2] attention mechanism. They use object properties in particular as attention candidates rather than focusing on the entire image divided into cells. These properties are extracted using a detector like Faster R-CNN [3,5] trained on the Visual Genome dataset. This approach marked a significant development for the VQA community and dramatically enhanced VQA performance [4]. The authors mix their developed modules with potential bottom-up object candidates to achieve cutting-edge performance.

Table 1. Accuracy on VQA Dataset

QA MODELS	TEST-DEV
VQA (Antol et al. 2015) [30]	53.74
Comp_QA (Andreas et al. 2015) [16]	54.8
MCB (Fukui et al. 2016) [34]	66.7
SAN (Yang al. 2016) [40]	58.7
SMem-VQA (Xu and Saenko 2016) [38]	58
Region_VQA (Shih et al. 2016) [31]	62.44
Ask_Neuron (Malinowski et al. 2017) [22]	58.4
SCMC (Cao et al. 2017) [12]	60.96
MLAN (Yu et al. 2017) [14]	65.3
CMF (Lao et al. 2018) [17]	66.4
IDA-NTN (Bai et al. 2018) [6]	67.56
MCAN (Yu et al. 2019) [11]	70.63
cycle (Shah et al. 2019) [39]	69.87
AnswerAll (Shrestha et al. 2019) [19]	65.96
Dropout (Fang et al. 2019) [26]	66.92

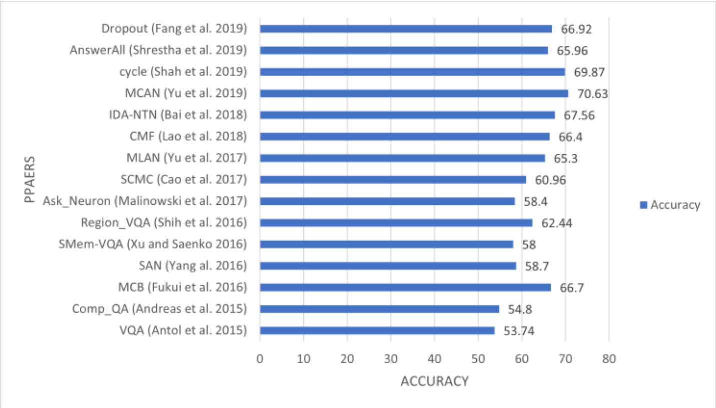

Fig 2. Accuracies on test-dev and test-std splits of VQA dataset

The findings of a research conducted on the VQA dataset using the test-dev are then presented in Table 1 and Fig. 2. The VQA dataset, which consists of 265,016 images and open-ended questions on the visuals, is rather huge. To respond to these inquiries, one must possess knowledge of common sense, language, and eyesight. The maximum recorded accuracy on this is at around 70%, indicating room for improvement. There is still a discrepancy between VQA performance and human performance for unknown causes. In order to bridge the gap between machine and human intelligence in picture perception, this section indicates prospective future work needed at several elements of VQA.

III. Recent developments in Visual Question Answering

The cornerstone for overcoming the constraints of the majority of the current VQA methodologies is Deep Learning techniques. The Multi-World QA method and the Answer Type Prediction (ATP) method stand out as the two outliers. Naturally, there are additional non-DL methods that are employed as comparison methods for different datasets and methods. The question is usually included in Deep Learning-based VQA systems utilising one of the word embedding approaches, like Word2Vec, occasionally in conjunction with an RNN. The majority of them also utilise CNN to spot details in the photographs. iBOWIMG, Full-CNN, Ask Your Neurons (AYN)[6], Vis+LSTM, Dynamic Parameter Prediction (DPPnet), and other methods are among them. Where to Look (WTL)[7], Recurrent Spatial Attention (R-SA), Stacked Attention Networks (SAN), Hierarchical Coattention (CoAtt), Neural Module Networks (NMNs), and other DL-based systems are a few examples of those that use an attention mechanism. For two reasons, the existing research covered in this part has no direct bearing on the VQA-Med. The focus on the medical industry, which presents a distinct set of problems for this topic, is the first and most visible component. The sentences of the answers in VQA-Med are structured differently from those in other VQA datasets, such as Dataset for Question Answering on Realworld Images (DAQUAR) [8], Visual7W [9], Visual Madlibs [10], COCO-QA [10], Freestyle Multilingual Image Question Answering dataset (FM-IQA) [11], Visual Question Answering (VQA) [6], etc. Neural-Image-QA, an initial approach to the VQA challenge, was proposed by Malinowski and Fritz. An LSTM generates the response after receiving the query and CNN Image attributes. The VQA problem was first tackled as a classification task with the introduction of the VIS+LSTM model by Ren et al. Antol et al. popularized the task by creating the VQA v1.0 dataset, evaluation measure and annual challenge. The baseline LSTM Q + norm I used activations from the final hidden layer of VGGNet as picture features, and could achieve comparable results using bag-of-words instead of an RNN. To improve visual attention, the authors created the SAN architecture with multiple attention layers. The multi-class model in the final iteration used attention as proposed by Andreas et al. Their NMT model broke down questions into linguistic structures to choose a modular network for prediction. Lu et al. developed the first attention model considering both question and picture attributes (co-attention). Standard attention models were used on CNN features that correspond to uniform grids. Anderson et al. used a two-phase method with a pre-trained R-CNN to estimate attention distributions. Fukui et al. initially solved VQA using bilinear pooling, which considers complex interactions between modalities but results in a large number of parameters. MUTAN and MLB approximate bilinear pooling and use the pre-trained Skip-thoughts encoder for question representation. Kazemi and Elqursh achieved comparable results by only concatenating the modalities, raising questions about the necessity of bilinear pooling. The 2021 VQA challenge was won by FAIR 3's Pythia v0.1, a reimagined bottom-up top-down approach. In medical VQA, knowledge is required to understand medical pictures, making it a specialized task. The first edition, VQA-Med 2018, had 6,413 question-answer pairs for 2,866 medical images. Deep learning was used by majority of the submissions, which employed pre-trained CNNs, LSTM/Bi-LSTM encoder-decoder architecture and advanced techniques like stacking attention networks and MCB pooling.

Using VGG16, a team from Jordan University of Science and Technology retrieved image attributes. They used an LSTM-based encoder-decoder model, sending the query to the encoder first and then the image attributes and the beginning hidden states of the decoder. The Faculty of Sciences and Techniques, Tangier, Morocco team tackled the problem as a multi-label classification issue. They used VGG16 to extract image features while also sending the word-embedded question into a Bi-LSTM network to extract question attributes. The properties of the combined query and image were then sent to a decision tree classifier. The TU team provided two models. They used the pre-trained Inception-ResNet-v2model and Bi-LSTM rather than LSTM to extract image features in the first model, which has roughly the same architecture. Their second model estimated the attention between the image features and the question features, which was then concatenated with the question features and passed to a Softmax layer for prediction.

Two further models were created by the National Library of Medicine, Bethesda, Maryland, USA team. The initial model used a stacked attention network (SAN) with VGG16 for picture features and LSTM for question features. They used Multimodal Compact Bilinear pooling (MCB) for the second model, which included 2-layer LSTM question features and ResNet-50 and ResNet-152 for image features. The second attention layer in the SAN model combines the image features and the query features, and after computing attention over the image, it sends the problem to the Softmax layer as a classification problem. Using external medical images, ResNet-50 and ResNet-152 were modified for the MCB model, and the image features and question features were then combined to create a multimodal representation to predict. The University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland team extracted picture features from ResNet-152 and used a pre-trained word embedding on Wikipedia pages, PubMed articles, and Pittsburgh clinical notes for text features. They created multiple attention maps by utilising a co-attention mechanism between textual and visual components. Then, using a sampling strategy, students provided responses as part of a categorization task.

A. General VQA Model

VQA models can be broken down into a variety of parts. To evaluate the visual data and retrieve the picture attributes, a CNN is typically utilized. An RNN and an embedding layer are used to construct a text encoder from the text input to provide a word vector of the query. Depending on the inquiry, an attention strategy may be used to concentrate on particular features of an image. In order to anticipate the correct response, a multi-layer classifier is given the attended visual input as well as the question representation.

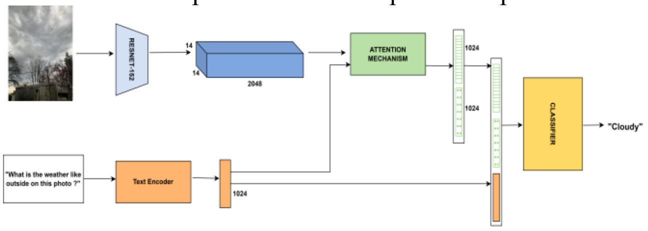

Fig 3. VQA Model

The images undergo preprocessing before being fed into ResNet-152. The images are resized and center-cropped to a uniform size of 448 x 448, then the data from each channel is normalized by subtracting the mean and dividing by the standard deviation. When deciding whether to use an attention mechanism, it's important to choose the right layer for attribute extraction. If not using attention, it's crucial to extract the activations from the penultimate layer of the pre-trained ResNet-152. This results in a 2048 x 2048 feature vector for each image.

B. Embedding Layer

The process of compressing semantic information relevant to the task of VQA into compact representations is called word embedding. A look-up table with a defined vocabulary and size for embeddings is added to the model. These embeddings can be pre-trained or generated randomly and fine-tuned during training. The performance of the VQA model was compared using pre-trained and scratch-trained embeddings. The Embedding layer module takes a list of index inputs and outputs corresponding word embeddings.

C. Recurrent Neural Network

Each query is approached by RNNs as a collection of word embeddings, and each embedding is repeatedly combined with its corresponding prior memory to produce a semantic representation of the entire phrase. Thus, utilising RNNs, the entire series, which has a variable length, may be included into an embedding with a fixed size. We compared the performances of a model using pre-learned RNN parameters with a model in which the parameters are randomly initialised when creating our final solution. The RNN parameters are tuned during the training of the VQA model.

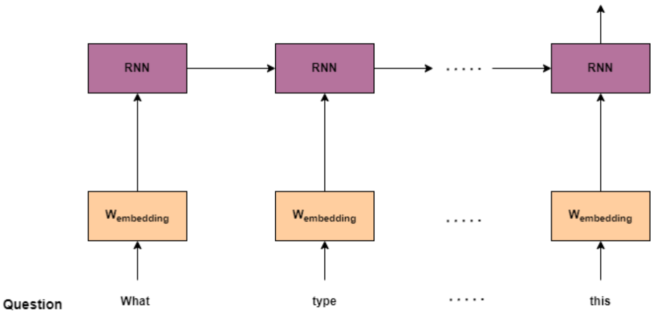

Fig. 4 Text encoder

D. Randomly initialized Embedding layer + LSTM

In the VQA model, we use an embedding layer with random initialization, so the word embeddings are learned during optimization of the VQA objective. The Text Encoder uses an LSTM RNN module, also trained from scratch. A text encoder that uses an LSTM (Long Short-Term Memory) RNN (Recurrent Neural Network) model is a popular approach for natural language processing tasks such as language translation, sentiment analysis, and text classification. The LSTM RNN model is particularly effective at processing sequential data, such as text, and capturing long-term dependencies.

The text encoder works by first tokenizing the input text into a sequence of words or subword units. Each token is then represented as a dense vector, typically through an embedding layer that maps each token to a low-dimensional continuous space. The embedding layer helps to capture semantic information about the tokens and enables the model to generalize better to unseen text.

The sequence of embedded tokens is then fed into the LSTM RNN model, which processes the input sequence one token at a time and updates its hidden state based on the current token and the previous hidden state. The LSTM's gated structure allows it to selectively store and discard information over time, which helps to capture long-term dependencies in the input sequence.

At each time step, the LSTM RNN model outputs a hidden state vector that represents the current state of the input sequence. This hidden state vector can be used as a fixed-length encoding of the input text for downstream tasks, such as classification or generation.

Overall, the text encoder that uses an LSTM RNN model is a powerful tool for processing sequential text data and capturing its underlying structure. By incorporating the strengths of the LSTM RNN model, the encoder can effectively encode text into a compact and meaningful representation that can be used for various natural language processing tasks.

IV. APPLICATIONS OF VISUAL QUESTION ANSWERING

Recent articles [19], [28], [30], and [3] have all discussed visual question answering. For instance, [28] only considers issues whose solutions come from a defined, constrained environment of 894 item categories or 16 basic colours. Furthermore, [19] considers questions generated from templates based on a predetermined vocabulary of objects, traits, linkages between things, etc. The following approaches tie the proposed VQA effort to related work: [30] studied the cooperative parsing of videos and the associated text in order to reply to queries on two datasets, each including 15 video clips. [3] uses crowdsourcing to find people to answer questions from users who are blind or visually impaired about visual information. In similar work by [21], a CNN for the picture and an LSTM for the question were coupled to generate a response. Their model, which bases the LSTM question representation on the CNN image features at each time step, use the final LSTM hidden state to progressively decode the response phrase. To produce a softmax distribution over output answer classes, the model developed in this study experiments with "late fusion," where the CNN image features and the LSTM question representation are computed separately, combined using element-wise multiplication, and then passed through fully connected layers. [4] creates abstract scenes to help people catch the visual common sense needed to reply to (purely text-based) fill-in-the-blank and visual paraphrase questions. Using visual evidence, [22] and [18] assess the plausibility of assertions made using common sense. In parallel with our study, [18] compiled queries and responses in COCO pictures. Using COCO, [17] automatically produced four different sorts of questions.

Text Based Question and Answering:
NLP and text processing have received considerable attention in recent research, with examples such as [15] [14][6] [5][41]. Sentence completion is another similar text task (e.g. [2] with multiple-choice answers). The methods used for VQA are influenced by these techniques. One of the key challenges in VQA is how the queries are derived from text, including [29], which used QA pairs with textual descriptions and simulated persons and objects in predefined locations. VQA requires a combination of both text (questions) and graphics as it is inherently visual (images). Having a good understanding of the text and advanced reasoning skills is especially important as people ask questions.

Describing Visual Content:
In VQA, generating words or sentences to describe visual content is a crucial step, as seen in tasks such as image labeling [11], [29], image captioning [30], [17], [8, [9], [16], [9], [12], [24], [3], [26], and video captioning [6], [21]. These tasks demand both visual and semantic proficiency, despite the often ambiguous nature of subtitles (as noted by [23]). Generic visual descriptions are not sufficient as VQA questions demand detailed and accurate information about the image.

Other Vision+Language Tasks:
Recent tasks that blend vision and language, such as the creation of referring expressions [25, 26] or determining coreference [28, 27] for specific objects in an image, are easier to evaluate than tasks like image captioning (e.g. "the one in a red shirt" or "the dog on the left"). Although referring expressions are task-specific and targeted, they often only involve a limited number of visual cues (such as color and position). Our work shows that visual questions and answers result in a more diverse range of visual concepts.

A. Medical Visual Question Answering(VQA) Model

An illustration of a typical visual question-answering technique that requires a large amount of labelled data to train is shown in Figure 7. Unfortunately, such vast amounts of data are rarely available to the medical sector. In this study, we present a innovative medical VQA paradigm that gets around the restriction of labelled data. The proposed framework examines the use of the unsupervised Denoising Auto-Encoder and the supervised Meta-Learning. DAE has the advantage of using a lot of unlabeled images, whereas meta-learning offers the benefit of experience meta-weights that quickly adapt to VQA issues with minimal labelled data. It is feasible to train the suggested framework well with a limited number of labelled training instances by utilising the benefits of these strategies.

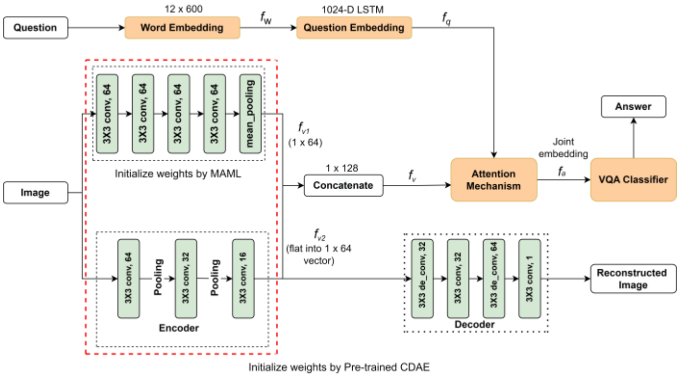

Fig 5. Medical VQA architecture

The sample conditions listed below were applied:
(i) To ensure that each image depicted a distinct patient, just one image was utilised for each instructional situation.
(ii) Each image is unique enough to show separate structures.
(iii) The images do not contain any observable radiological indicators, such as dots or circles.
(iv) Image captions adequately describe at least one structure and match the image.

Helping physicians with radiography involves a variety of tasks, from assisting those who are just studying the fundamentals of reading a plain film x-ray to assisting experienced radiologists who are looking at unique sickness presentations. We only selected medical students and fellows who have at least completed their core clinical rotations since they will have a better understanding of how radiography is integrated into regular clinical decision making and patient care. The questions and assignments provided to these trainees are more similar to those given to visiting trainees than they are to first-year medical students. We do predict that resident trainees will have higher expectations as they become more skilled at interpreting radiological images and formulating more difficult questions.

B. VQA in Advertisement

The process of compressing semantic information relevant to the task of VQA into compact representations is called word embedding. A look-up table with a defined vocabulary and size for embeddings is added to the model. These embeddings can be pre-trained or generated randomly and fine-tuned during training. The performance of the VQA model was compared using pre-trained and scratch-trained embeddings. The Embedding layer module takes a list of index inputs and outputs corresponding word embeddings.
1. ViSe Network:
A potential action-reason statement and an image are each inputs to separate visual and semantic properties in this network, which is made up of many visual and semantic streams. The concatenated "action-reason" statement is utilised as the typical "question" phrase used as input in a VQA formulation, and it builds on the visual-semantic VQA framework that took first place in the 2017 VQA competition. By applying bottom-up attention, the visual

stream separates distinct visual elements. Each word in the input sentence is looked up and processed by the semantic stream. The network integrates high level visual and semantic information to produce two output probabilities for the 0 and 1 classes.

2. SymViSe Network:
The ViSe Network, which outputs symbolic attributes from the input picture using a second deep visual network built similarly to the one used in the visual stream, is combined with a third symbolic stream in this network, as illustrated in Figure 8. Using the symbolic annotations from the Visual Advertisement dataset, this visual encoding network has been adjusted to be able to infer symbolic meaning from items in the image. The visual and semantic data gathered from the other streams are projected into a joint-embedding space along with these symbolic properties. To calculate the degree of agreement between the input image and the text, the visual and symbolic streams are multiplied by the semantic stream, joined, and then passed through a linear fully-connected classifier.

3. Visual Stream:
The components in a visual advertising might be used to represent it. We pre-train a Faster R-CNN model to recognise the 80 items in the COCO dataset because there are no objects that have labels in the advertising dataset. As an image encoder, create K = 100 feature vectors with 1024 dimensions using this model. To facilitate cross-modal transfer learning, these vectors are L2 normalised and scaled into a predictable range.

4. Semantic Stream:
Figure 8 illustrates our decision to encode each action justification phrase as a 20x300 feature vector made up of concatenated GloVe embeddings for each word. Sentences under 20 words are zero-padded to keep the word count consistent. Only the first 20 words of phrases longer than 20 words are utilised. These textual features are encoded using a two-layer LSTM network and then mapped onto a joint using a fully connected non-linear layer.

5. Symbolic Stream:
The underlying abstract semantic ideas cannot be adequately communicated using the elements of visual advertising. The symbolic significance of things and the significance of their placement within the composition are concepts we want to instil in our model. We train a Faster R-CNN model to extract picture regions corresponding to the 1000 most frequent symbols in the Visual Advertisement dataset. It preprocesses the free-response symbol annotations before deciding on the candidate symbol classes. It will try to use an auto-correct tool to fix misspelt symbol names. In order to distinguish between various symbol conjugations, words are lemmatized. Before calculating the words' tf/idf scores, we map each word in the training set to its correct stem in order to count the total occurrences of each at symbol. From the training set, we choose the top 1000 words to use as the candidate set. We choose the symbol with the greatest tf/idf score as the associated symbol for each bounding box because we want to be able to name each box uniquely. The bounding boxes that can't be assigned to a legitimate symbol class. The ground-truth symbolic bounding box annotations from the ads dataset are used to individually train the model. This job is more challenging than basic object recognition since items with a wide variety of exterior attributes are assigned "fun" and "hazardous" symbol labels. The R-CNN creates K = 100 symbolic feature vectors with 1024 dimensions for the ViSe and SymViSe models, which are then L2 normalised to scale into a common range to facilitate cross-modal transfer learning.

6. Bottom-Up Attention for Visual/Symbolic Feature Extraction:
Extraction of activity maps from a model trained on a fundamental task, such as image classification, is a com`mon technique for modelling visual properties. This technique maintains representations of the background class, but it may also favour characteristics that are less representative of the image's key parts. When seeing a scene, a person may consciously utilise high level attention to concentrate on particular features or they may be drawn to particular locations based on low level qualities. By removing properties from an image that correspond to the things in it, bottom-up attention makes an attempt to represent this subsequent effect. In order to encode picture attributes in the visual and symbolic streams, we employ the bottom-up attention paradigm. To do this, we modify a Faster-RCNN model using a 101-layer ResNet feature extractor. According to the model's assumption that an object or symbolic property exists, it ranks bounding box predictions. We mean-pool the feature vectors in the region that the matching box on the up-sampled activity map covers for each of the top K predictions. Thus, a single feature vector that is associated to each prediction represents the semantic content of the prominent region. It makes use of extraction modules designed specifically to find various types of semantic material.

Despite all the advancements in recent years, the majority of these techniques still have some potential limitations. The first issue with present methodologies is how to respond to a question that necessitates a lengthy chain of deductions. Additionally, it appears that none of these systems are very adept at handling queries requiring quick recall, such those involving integer equality. A query about counting the number of certain items in the image is another illustration of a problem that might be particularly challenging for the majority of the models we discussed in this study. There have recently been initiatives to solve these difficulties as well. The things that contribute to each count are also identified by this reinforcement learning method. Future models can enhance the effectiveness of present techniques by expanding on them, such as co-attention or modular networks, while also addressing the issues raised above, perhaps by employing a solution specifically designed to handle them. The issue with this work is that it only employs synthetic pictures in the CLEVR dataset, despite the fact that the technique appears to be promising. It could be interesting to work on creating a real-world implementation of LBA (learning-by-asking) in the future by swapping out the synthetic photos with actual situations.

Visual Question Answering (VQA) is a challenging research problem that requires an algorithm to answer questions based on the visual content of an image. Despite significant progress in recent years, there are still several challenges that need to be addressed in VQA research. Below are some suggestions and opportunities to overcome these challenges:

1. Limited Generalization: VQA models often struggle to generalize to new or unseen images and questions. To address this challenge, researchers can explore transfer learning techniques, such as pre-training models on large-scale image and language datasets. Fine-tuning the pre-trained models on smaller VQA datasets can also improve generalization.
2. Ambiguity: Images and language are inherently ambiguous, which can make it difficult for VQA models to accurately answer questions. To overcome this challenge, researchers can explore incorporating common sense knowledge or external information sources into VQA models.
3. Biases: VQA datasets can contain biases that can affect model performance. Researchers can use debiasing techniques, such as data augmentation or bias-aware training, to mitigate these biases.
4. Limited Understanding of Images: VQA models often have limited understanding of the visual content of an image, which can make it difficult to accurately answer questions. Researchers can explore multi-modal approaches, such as combining image and text features, to improve understanding of the visual content.
5. Limited Understanding of Language: VQA models may also struggle to understand complex language, such as idiomatic expressions or sarcasm. To address this challenge, researchers can explore incorporating contextual information, such as discourse analysis, into VQA models.
6. Lack of Diverse Datasets: VQA datasets often lack diversity, which can limit model generalization and lead to biases. To overcome this challenge, researchers can collect more diverse datasets or use data augmentation techniques to increase dataset diversity.

Overall, the key to overcoming these challenges is to explore a variety of techniques and approaches, including transfer learning, multi-modal approaches, and incorporating external knowledge sources. Additionally, it is essential to continuously evaluate and improve VQA models to ensure they are robust and reliable.

Conclusion and future work

To be significantly more capable than task-specific techniques, such as object recognition and object detection, VQA is a key research issue in machine learning and computer vision. The ability to complete VQA tasks that can respond to any image-related enquiry is an advance in artificial intelligence. The datasets and techniques for VQA that are currently offered in this survey have undergone a thorough evaluation. This review covers a number of problems that biases and other problems cause with existing datasets. Larger and more varied datasets for VQA will continue to be developed in the future. VQA algorithms that can analyse visual content will require more work to develop, but they may also lead to the development of significant new academic fields. Building efficient, impartial, and goal-oriented datasets for assessing important features of VQA should be the main focus of emerging VQA research. In order to generate more effective multi-modal fusion—a job for which joint attention mechanism has proven to be beneficial—Deep CNN may be utilised to extract more natural visual aspects and mix them with the most recent word embedding models.

References

[1] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu. Advanced deep-learning techniques for salient and category specific object detection: a survey. IEEE Signal Processing Magazine, 35(1):84–100, 2017.

[2] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6077-6086).

[3] S. R. Kheradpisheh, M. Ganjtabesh, S. J. Thorpe, and T. Masquelier. Stdp-based spiking deep convolutional neural networks for object recognition. Neural Networks, 99:56–67, 2017.

[4] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2017.

[5] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy. Improved image captioning via policy gradient optimization of spider. In Proceedings of the IEEE international conference on computer vision, pages 873–881, 2017.

[6] Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271.

[7] Wu, Z., Palmer, M.: Verbs semantics and lexical selection. In: Proceedings of the 32nd annual meeting on Association for Computational Linguistics. pp. 133–138. Association for Computational Linguistics (2020)

[8] Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 21–29 (2020)

[9] Yu, L., Park, E., Berg, A.C., Berg, T.L.: Visual madlibs: Fill in the blank description generation and question answering. In: Computer Vision (ICCV), 2015 IEEE International Conference on. pp. 2461–2469. IEEE (2020)

[10] Zhou, B., Tian, Y., Sukhbaatar, S., Szlam, A., Fergus, R.: Simple baseline for visual question answering. arXiv preprint arXiv:1512.02167 (2020)

[11] Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., ... & Bolton, E. E. (2019). PubChem 2019 update: improved access to chemical data. Nucleic acids research, 47(D1), D1102-D1109.

[12] Zhao, Y., Tang, S., Guo, J., Alahdal, M., Cao, S., Yang, Z. & Jin, L. (2017). Targeted delivery of doxorubicin by nano-loaded mesenchymal stem cells for lung melanoma metastases therapy. Scientific reports, 7(1), 1-12.

[13] X. Chen and C. L. Zitnick. Mind's Eye: A Recurrent Visual Representation for Image Caption Generation. In CVPR, 2015.

[14] Yu, D., Fu, J., Mei, T., & Rui, Y. (2017). Multi-level attention networks for visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4709-4717).

[15] 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.

[16] Luketina, J., Nardelli, N., Farquhar, G., Foerster, J., Andreas, J., Grefenstette, E., ... & Rocktäschel, T. (2019). A survey of reinforcement learning informed by natural language. arXiv preprint arXiv:1906.03926.

[17] Lao, M., Guo, Y., Wang, H., & Zhang, X. (2018). Cross-modal multistep fusion network with co-attention for visual question answering. IEEE Access, 6, 31516-31524.

[18] 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.

[19] Shrestha, M. Hejrati, A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every Picture Tells a Story: Generating Sentences for Images. In ECCV, 2019.

[20] D. Geman, S. Geman, N. Hallonquist, and L. Younes. A Visual Turing Test for Computer Vision Systems. In PNAS, 2014.

[21] Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., & Parikh, D. (2017). Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6904-6913).

[22] Kafle, K., & Kanan, C. (2017). Visual question answering: Datasets, algorithms, and future challenges. Computer Vision and Image Understanding, 163, 3-20.

[23] S. Kazemzadeh, V. Ordonez, M. Matten, and T. L. Berg. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In EMNLP, 2014.

[24] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. TACL, 2015.

[25] Barra, S., Bisogni, C., De Marsico, M., & Ricciardi, S. (2021). Visual question answering: Which investigated applications?. Pattern Recognition Letters, 151, 325-331.

[26] Fang, Z., Liu, J., Li, Y., Qiao, Y., & Lu, H. (2019). Improving visual question answering using dropout and enhanced question encoder. Pattern Recognition, 90, 404-414.

[27] Gupta, A. K. (2017). Survey of visual question answering: Datasets and techniques. arXiv preprint arXiv:1705.03865.

[28] X. Lin and D. Parikh. Don't just listen, use your imagination: Leveraging visual common sense for non-visual tasks. In CVPR, 2015.

[29] Ren, F., & Zhou, Y. (2020). Cgmvqa: A new classification and generative model for medical visual question answering. IEEE Access, 8, 50626-50636.

[30] 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).

[31] Shih, K. J., Singh, S., & Hoiem, D. (2016). Where to look: Focus regions for visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4613-4621).

[32] V. Ramanathan, A. Joulin, P. Liang, and L. Fei-Fei. Linking People with "Their" Names using Coreference Resolution. In ECCV, 2014.

[33] M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. In NIPS, 2015.

[34] Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., & Rohrbach, M. (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847.

[35] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele. Translating Video Content to Natural Language Descriptions. In ICCV, 2013.

[36] Alberti, C., Ling, J., Collins, M., & Reitter, D. (2019). Fusion of detected objects in text for visual question answering. arXiv preprint arXiv:1908.05054.

[37] K. Tu, M. Meng, M. W. Lee, T. E. Choe, and S. C. Zhu. Joint Video and Text Parsing for Understanding Events and Answering Queries. IEEE MultiMedia, 2014.

[38] Xu, H., & Saenko, K. (2016). Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14 (pp. 451-466). Springer International Publishing.

[39] Shah, M., Chen, X., Rohrbach, M., & Parikh, D. (2019). Cycle-consistency for robust visual question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6649-6658).

[40] Hajissa, K., Islam, M. A., Sanyang, A. M., & Mohamed, Z. (2022). Prevalence of intestinal protozoan parasites among school children in africa: A systematic review and meta-analysis. PLoS neglected tropical diseases, 16(2), e0009971

[41] Bathla, G., Bhadane, K., Singh, R. K., Kumar, R., Aluvalu, R., Krishnamurthi, R., ... & Basheer, S. (2022). Autonomous vehicles and intelligent automation: Applications, challenges, and opportunities. Mobile Information Systems.