DOI: 10.15649/2346030X.3370
Aibi revista de investigación, administración e ingeniería. Volume 11, Number 3, Pág 126-142, September - December 2023, ISSN 2346-030X

A comprehensive review and open challenges on visual question answering models

Dipali Koshti[1], Ashutosh Gupta[2], Mukesh Kalla[3], Arvind Sharma[4]
[1,2,3,4]*Sir Padampat Singhania University, Rajasthan - India*
ORCID: [1]*0000-0002-8642-8002*, [2]*0000-0003-3257-0836*, [3]*0000-0002-6981-0963*, [4]*0009-0000-1234-4321*

Received: July 05, 2023.
Accepted: August 15, 2023.
Published: September 01, 2023.

***Abstract*—** - Users are now able to actively interact with images and pose different questions based on images, thanks to recent developments in artificial intelligence. In turn, a response in a natural language answer is expected. The study discusses a variety of datasets that can be used to examine applications for visual question-answering (VQA), as well as their advantages and disadvantages. Four different forms of VQA models—simple joint embedding-based models, attention-based models, knowledge-incorporated models, and domain-specific VQA models—are in-depth examined in this article. We also critically assess the drawbacks and future possibilities of all current state-of-the-art (SoTa), end-to-end VQA models. Finally, we present the directions and guidelines for further development of the VQA models.

Keywords: VQA review, Image-question answering, Visual question answering.

*Corresponding author.
Email: dipalis@fragnel.edu.in (Dipali Koshti).
Peer reviewing is a responsability of the Universidad de Santander.

How to cite this article: D. Koshti1, A. Gupta, M. Kalla and A. Sharma, "A comprehensive review and open challenges on visual question answering models", *Aibi research, management and engineering journal*, vol. 11, no. 3, pp. 126-142 2023, doi: 10.15649/2346030X.3370

I. INTRODUCTION

Rapid developments in the realm of Deep Learning have opened the doors for many new applications. Many of the applications we use in our day-to-day lives are multi-disciplinary, meaning integrating ideas from multiple disciplines to solve a real-world problem. Visual Question Answering is one such application that has drawn a lot of interest from several research communities, particularly those in image and text processing. The multidisciplinary field of vision question answering (VQA) calls for expertise in three different fields: knowledge reasoning, image processing, and language processing. VQA is an intelligent technology that allows us to enter an image and a pertinent query. The machine outputs the answer, which is likewise in natural language sentence form. Thus, this problem by nature is multi-disciplinary. It demands Computer vision skills since we need to understand the image content in order to perform many image-related tasks such as detecting scenes, counting objects, detecting Objects, detecting colors, etc. Also, VQA demands the knowledge of language Processing (NLP) as the system needs to process the question in order to understand its semantic information and context. And, finally, we need knowledge reasoning as some of the questions may require knowledge from outside the image. Sometimes we need to extract external knowledge from the knowledge bases and combine extracted knowledge in a system to answer the question.

The origin of VQA has come from Facebook's Mr. Matt, who himself is a blind person. Facebook designed an automatic caption generation system for photos that are uploaded on Facebook and these captions can be read aloud for blind users. This idea is then extended to Visual Question Answering, which will in the future help blind people to answer any question about the image they ask. Later, this concept was extended to other applications such as answering questions about medical images, Satellite images, and even answering questions from data visualization.

While such a kind of visual reasoning is easy for humans, it's a bit of a stimulating task for a machine. VQA is challenging since the questions have to be answered from the images. Users may ask any open-domain questions about the image in natural language. The answers generated may be one word or phrase containing a few words. In general, any VQA system includes mainly three phases shown in Figure 1:

Phase 1: Extraction of question and image features.
Phase 2: Combining question features with image features for joint comprehension.
Phase 3: Generation of answer.

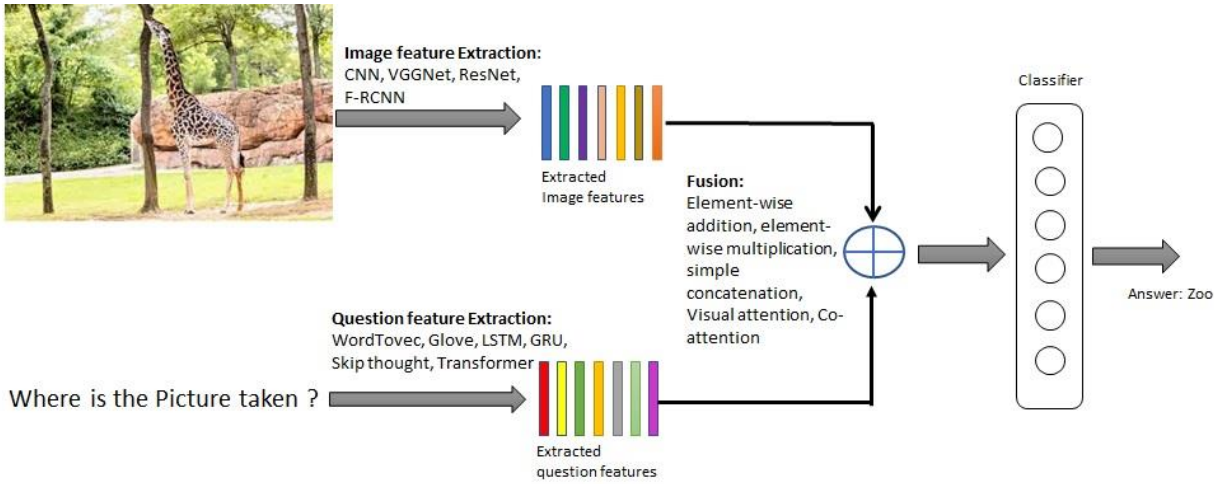

Figure 1: Visual question answering - Three phases.
Source: Own elaboration.

A variety of techniques have been investigated in the past for each of the aforementioned phases. The most popular method for phase one is to extricate picture features using CNN and their derivatives, such as GoogleNet ResNet, VGGNet, and Recurrent Neural Networks and their variations like LSTM and GRU are also accessible for question feature extraction. Phase 2 is a little difficult since it fuses the image features and the question elements to understand how they relate to one another. The literature has examined a variety of techniques, including basic concatenation, element-wise multiplication, and complicated attention networks.

The contribution of this study is as follows.

1 The study discusses various datasets available for exploring VQA applications along with their limitations.
2 The paper presents a comprehensive review of four categories of VQA models: Simple joint embedding-based VQA models, Attention-based VQA models, Knowledge Incorporated VQA models, and domain-specific VQA models.
3 We critically analyze all the state-of-the-art (SoTa), end-to-end models, for VQA, their limitations, and future improvements.
4 Finally, we provide the guidelines and future direction for further improvements in VQA models.

The contents are arranged as follows. Section 2 discusses details of image featurization and questions featurization methods. Section 3 discusses various fusion techniques used for combining image features and question features, and section 4 talks about various methods employed for answer generation. Section 5 focuses on Knowledge incorporated VQA models, and section 6 discusses domain-specific VQA models. A variety of VQA datasets have been discussed in Section 7. Various evaluation metrics used to evaluate VQA models, their merits, and demerits have been discussed in section 8 and finally, we deduce our findings of the study in section 9.

II. FEATURE EXTRACTION

As discussed earlier, the first phase in the VQA model is to extricate visual features and question features. Visual feature extraction involves extracting important features from a query image and representing the image in its numeric form so that it can be further processed by neural networks. Most VQA literature utilizes Convolution Neural Networks (CNN) and their variants for image featurization. In the beginning, many researchers used VGGNet for the extraction of image features [1, 2, 3, 4, 5]. They used the final hidden layer of VGG-Net as image features as most of the spatial information is retained in the last hidden layer. Sometimes, attributes are taken out from the last pooling layer instead of the last inner product layer [6]. Also, ResNet is found to be one of the demanded networks used for image feature extraction after its inception. [7,8,9] have used ResNet for image feature extraction. One of the reasons why VGGnet is still preferred over ResNet is because of its simple and lightweight architecture and fast convergence compared to ResNet which is four times heavier than VGGNet. But with the availability of high computing resources now it is possible to train the model using different ResNet networks like ResNet18, RestNet108, ResNet152 etc. [7,8,9,10].

Global aspects of the image were employed in the aforementioned publications. Some research has shown that capturing regional or local information enhances model performance even more. Faster R-CNN was utilized by [9, 11, 12, 13] to extract local object-level characteristics from the image. Local features make it possible to explore the image's more detailed and important features. Ilievski et al. retrieved features from those objects in [5] that are associated with the keywords asked for. To extract object and picture features, they employed ResNet. They investigated the image's object-level and image-level elements, fusing them to produce visual features.

Most of the literature after 2018 used transformers for image and language modeling since transformers have set a new benchmark for most of the vision + language tasks. [13] proposed a complete BERT-based model where two separate BERT-style models were used for image and language features. LXMERT [14] constructed a fully transformer-based VQA model where they used BERT not only for language modeling but also for vision modeling. The completely BERT model outperforms previous transformer-based models, according to researchers who built three transformer encoders—a question encoder, an object connection encoder, and a cross-modality encoder—and pre-trained the model on five distinct cross-modality tasks. UNITER [15] used conditional masking for pretraining tasks and proposed a novel method for exact alignment between word regions and image regions. In [16], writers looked into object tags as anchor points for text and obvious visual object alignment in paired questions. They employed multi-layer converters built on BERT to incorporate images and quizzes. Table 1 summarizes deep learning models used in the past for image feature extraction.

Table 1: Different Models for Image feature extraction.

Method	**Paper**
VGGNet	LSTM+Q+I[1], AVWAN[2],SAN[6],Facts-VQA[70],DPP[75], QAM[76], Region-Sel[77], NMN[78],
ResNet	FDA[5], Bayesian [7], Dense-Sym[8], Code-Mix VQA[9], Hei-Co-atten [10], Rich-img-Region[27], MCB[29], MRN[30] , FVTA [33], MUTAN [36], Meta-VQA [77],Rich-VQA [79], QTA[80], , DCN [81],
GoogleNet	Neural Image QA [80], Multi-Modal QA [82] , i-Bowing[83], Smem[84]
F-RCNN	Code-Mixed VQA [9], CAQT [11], QLOB [12], BAN[28] , MFB[32], [85] ,explicit-know-Based[86] , Know-Base Graph[87]
BERT	VilBERT [13], LXMERT [14] , UNITER [15], Oscar [16], MPC [25], Semantic VLBERT [88]

Source: Own elaboration.

The next step in the VQA model is to extract question features. Before question features are extracted, word embedding process is performed. A number of word embedding techniques have been used in NLP. Some of these are one-hot encoding, Continuous bag of words (CBOW), co-occurrence matrix, word2vec, etc. Although the above word embedding techniques extract the contextual information from a given text, new recurrent networks such as LSTM and GRU have proven better to extract the contextual meaning of the question. But these networks don't exist independently. The basic idea is to create a word vector using any of the word embedding methods discussed above (e.g. word2vec) and these vectors are then fed to LSTM or GRU network.

[1,17] used a BoW (Bag of Words) approach. They made use of the fact that the words used to begin a question and an answer have a significant link. To provide solutions to multiple-choice question answering, [18] used word2Vec to extract question features. For each question-answer pair, fixed-length vectors were constructed and Stanford parsers were used to create four semantic bins. Bin 1 represents the question's type, Bin 2 represents its subject, Bin 3 represents all other noun terms' means, and Bin 4 represents all other words' means. After then, a bin for the terms in the candidate's response was created by concatenating the contents of all four bins. [19] KAN uses Golve embedding plus LSTM for question featurization as well as for extracted knowledge featurization. The language model utilized by the authors in [32] was glove plus GRU. [20] came up with a PRS structure to organize the data in a question, where P is the primary object, R is the relation, and S represents the secondary object. This method solved the binary visual question answering problem. Ideal values for P, S, and R would be nouns, verbs, or prepositions. LSTM and word2Vec were utilized for question feature engineering. One hot vector for word embedding, followed by an LSTM network, was utilized by [3] and [6] to extract semantic information from a query. Without utilizing LSTM, [7] investigated skip-thought word embedding algorithms. For encoding a question, writers in [21] employed a Skip-thought recurrent model that was trained on the Book corpus dataset. Although book corpus is from different domains, it works for remote sensing, thus training a model on different domains and applying it on some different models still works. [4,10] created an architecture that is hierarchical in nature that co-attends to the image and question at three different levels viz. the word, phrase, and question levels. They use an embedding matrix to embed words into a vector space at the word level. Unigrams, bigrams, and trigrams are utilized at the phrase level to store information using 1-D convolutional neural networks. They use recurrent neural networks to encode the complete question at the question level. In [11] authors used one hot encoding + Bi-LSTM. Instead of using the last hidden layer, they used all the hidden states of Bi-LSTM as the final features of the question. [23] used Glove for word embedding followed by GRU to extract question features.[16] used single directional GRU whereas [24] used Bi directional GRU for question embedding.

Recently transformers have gained popularity in NLP due to their capability of processing long sequences parallelly. Most of the recent work uses transformers for question featurization. [25,26] utilized transformer BERT for language feature extraction. In [87] authors concatenated the output of four consecutive BERT layers in order to generate hierarchical features from the question. [13] explored the power of transformers for both visual and language feature extraction. In [40] Alberto Mario et al. used OpenAI's GPT-2 Transformer as the language model. [27] used Transformer (BERT) for question feature extraction. They replaced the BANs [28] language model with BERT and used 20 Models of ensemble learning. Table 2 summarizes techniques used for question feature extraction.

Table 2: Word embedding techniques.

Method	Paper
One hot encoding	Ask your Neuron [3], FDA [5], CAQT[11]
BoW	VQA [1], Simple Baseline[17]
Glove	KAN [19],BAN[28] , TipsNTracks[89]
Word2Vec	Where- to-look [18],Yin And Yang[20]
Skip thought	Answer Type pred [7], RSVQA [59]
LSTM	CAQT[11], KAN [19], Yin and Yang[20], Know-Augmented VQA[46]
GRU	Coarse to fine reasoning [23], DMN [24], BAN [28], DPP [75]
Transformers	VilBERT[13],LXMERT[14], UNITER[15], Oscar[16], Coarse to fine reasoning [23], MPC [25], Hie-Alternation co-attention [26], Rich Image region VQA[27], KRISP[50]

Source: Own elaboration.

III. FUSION TECHNIQUES

To understand how the retrieved image and question features relate to one another, they must be combined. Information fusion is the process of combining two feature vectors—Visual features and Language features. Simple operations on feature vectors, such as concatenation, element-wise summing, or element-wise multiplication, can be used to do this [1,6,18]. Fukui and others [29] proposed MCB, a new joint representation technique that is easy to use but effective. It computes the outer product of two vectors, allowing all of the elements of both vectors to interact multiplicatively. Methods of attention-based fusion are used in recent publications. The network may focus more of its attention on key elements of the image and question, thanks to the attention mechanism. The attention-based mechanism uses one modality to determine the weights for the other modality. So, in the context of VQA, we can classify attention-based fusion into the following three categories:

1) Visual attention-based joint representation
2) Language attention-based joint representation
3) Co attention (Visual + Language) based joint representation

In a visual attention-based representation, the network only focuses on the image's central elements that are relevant to the query. Here question features are used to attend to key entities of the image. In language attention, we focus only on important words of the question sentence that are related to the image. In other words, important aspects of the question are highlighted using image attributes. Additionally, in co-attention fusion, visual features are utilized to attend question features while question features are used to attend visual features. [5,18,30, 31] used visual attention for their VQA model. In [5] the model learns the weights of visual features to select a small portion of the image. Model stacks the attention network for multiple reasoning thus narrowing down the selection region. Liani et al [31] proposed question-guided object attention; the model selects only question-related object features. In [10] authors introduced two types of co-attention models: 1) Parallel co-attention generates question and image attention simultaneously and 2) alternate-contention sequentially alternates between generating image and question attention. [2] introduced triple attention: Question attention, Mood attention, and Image attention. The model concatenated Question and mood features along with Image features. [32, 33, 34] all used a co-attention mechanism for the fusion of two modalities. In [35] Sehng Zeng focused on the features of the objects in an image and computes feature-wise attention weights for an object. They suggested a co-attention module that considers the features of related things, such as color, shape, and others apart from spatial information present in the object. MUTAN [36] presented a novel fusion method that explores the bilinear high-level interaction between question features and image features using tensor-based Tucker decomposition.

[37] proposed a model that creates a semantic feature vector for each (S,P,O) triplet where S is a Subject, P is a Predicate, and O is an Object in the image The semantic Relation parser generates relation triplets for each image in the form of SPO; the generated SPOs are then passed to the Semantic feature extraction model. Thus, visual features, Question features, and relation features are then fused together to generate mutual and self-attended representation. In [38] authors have extracted image features, image predicate features, question features, and question predicate features. These four feature vectors are then given to coarse-to fine-grained reasoning module where information filtering is done to filter out irrelevant information from image and question features. These filtered features are then forwarded to the multi-modal learning block that learns the semantic mapping between these features and finally, semantic reasoning is performed to generate an answer.

IV. ANSWER GENERATION

The final stage is to produce the pertinent answer after the joint representation of the picture characteristics and question features has been obtained. The literature has covered a variety of answer-generation approaches. The methods for producing replies are several and are covered below. 1) Open-ended, free-form inquiries where the responses could be in the form of single words, phrases, or even full sentences. 2) Questions that required you to count the number of things in an image in order to find the solution. 3) Multiple-choice tests. 4) Binary (yes/no) questions. A sigmoid layer is frequently used at the end of binary questions. One or two fully-connected layers are passed through with the joint representations. The output is sent through the classification layer, which is a single-neuron layer.

By choosing the top k most frequent answers in a dataset and predicting the top 5 most matched responses from these k answers, the majority of multiple choice VQA models have transformed the answer generation problem into a classification challenge. For VQA 2.0, the majority of researchers used k=1000. K=2000 and other values of k can also be investigated. MLP (Multi-layer Perceptron) has continued to be the method of choice for researchers to solve the VQA problem. [1,2,6,7,9]. The joint feature representations are often translated into replies for free-form, open-ended queries using a recurrent network like LSTMs. In their method, Malinowski et al. [3] combined the visual feature with an LSTM's representation of each word.

All of the above literature treated VQA as an answer selection (Discriminative) model. But in reality, VQA is a generation problem. Very few attempts have been made to take this challenge. [39] proposed a generative model for Med-VQA and also introduced a large-scale dataset called PMC-VQA for medical images. The model MedVInT was trained by aligning visual data from a pre-trained vision encoder with language models. In [40] authors unified both classification and generation in one model using a masked transformer. The model is capable of doing classification by using image and language features and generation by using picture, questions, and masked answers. Although the above two pieces of literature successfully converted VQA into a generation problem still there is a challenge when open-domain questions are posed and require external knowledge.

V. KNOWLEDGE INCORPORATED VQA MODELS

The literature mentioned above is capable of responding to closed-domain queries, that is, queries that can be resolved by examining the image and the query alone. However, in practice, we are free to pose any open-ended inquiry regarding the image that cannot be resolved by simply looking at the image and the question. To answer such questions, we may need to take into account knowledge beyond the visual concepts present in an image. Examples of closed-domain and open-domain questions are shown in Figure 2.

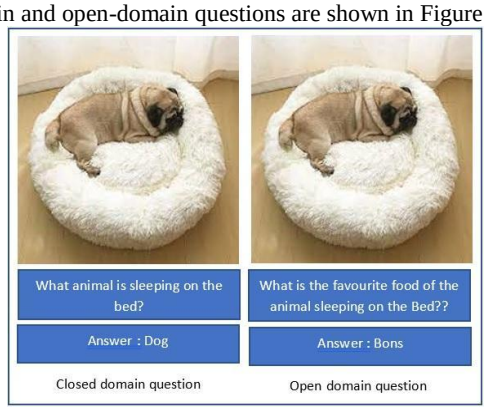

Figure 2: Example of Closed-domain and Open domain questions.
Source: Own elaboration.

A number of knowledge Bases (KBs) are available to explore knowledge incorporated VQA. Some of these KBs which are publicly available are Wikidata, DBPidea, ConcepNet, Webchild, and WordNet. Marino et. al [50] created a new benchmark, OK-VQA that contains most of the open domain questions where information provided in image and question alone is not enough to answer the question. Many authors have used a graph-based approach [17,41, 42,43] to integrate external KB into the VQA model where the important objects in the visual image represent a node in a graph and the relationship between these nodes as edge. [38] proposed a novel methodology that takes the question's textual keywords and important visual items from the image, using these to extract knowledge from ConceptNet in the form of a knowledge graph. Garderes et al. [44] Used ConceptNet as a Knowledge base. The model uses F-RCNN for extraction of the Image feature and BERT embedding for extraction of question features. They used Graph Convolution Neural network to extract knowledge from KB. [45] Extended conventional VQA dataset, called FVQA by adding supporting facts for each question-answer pair. The visual concepts from the image are extracted in the form of triplet <objects, scene, attributes> and represented in the form of a knowledge graph (KG) and linked to the extracted external knowledge from three different sources- DBPedia, ConceptNet, and Webchild. In [46] authors constructed scene graphs representing the object as nodes and relationships as edges. Based on the sentence similarity score the external knowledge has been extracted and constructed a scene graph which represents the relation of visual concepts with the objects as nodes and relationship as edges. After extracting the external information, the knowledge retrieval module computes sentence-level similarity scores and feeds the highest scoring knowledge entities to a Concept graph. The objects and knowledge instances were given weights using the Graph Attention Networks (GATs). Higher weights were assigned to the more relevant objects and knowledge instances with respect to the given question. Joint language, vision, and knowledge embeddings are represented by fusing the outputs of scene graph, concept graph, and question embedding. [47] basically created two graphs: the Scene graph and the external knowledge graph. The scene graph captures the relation between entities and predicates in an image. And the knowledge graph consists of general concepts and predicates acquired from external knowledge repositories. Both the graphs are then connected to generate the knowledge enriched final scene graph. This paper uses ConceptNet, WordNet and Visual Genome as KBs. Wu Qu et al. [48] developed a VQA model in which attributes of the image are extracted by CNN and five Image captions for each image are generated. The extracted attributes and generated captions are combined with the mined external knowledge (DBPedia) and are given to an LSTM to generate the answer. Shah et al. [49] created a new dataset called KVQA that included facts about persons. The model first identifies entities present in an image and question and then Identified persons (entities) are linked to the entity in WikiData. They used NER (Named Entity Recognition) to extract named entities from the question and related knowledge Facts are extracted from Wikidata. [50] combines the implicit knowledge acquired from question and image using a transformer (BERT) and explicit symbolic representation in the form of a knowledge Graph using four different knowledge sources DBPedia, ConceptNet, Visual Genome, and Haspart KB. [51] created their own dataset called Text-KVQA containing three categories of images: Scene, books, and Movies. They constructed knowledge bases for three domains: KB – books, KB – Movies, and KB- Business. These KBs have been constructed from three publicly available knowledge sources: Wikidata, IMDb, and book catalog. They explored the text present in the image. Textual features, Visual features, and question features were used to extract facts from KBs. [52] used all three KBs - WebChild, ConceptNet, and DBpedia to extract a total of 193, 449 facts. Each fact contains a triple: < subject, relationship, object>. Two modules—the sub-graph extraction module and the sub-graph embedding module—were employed in the model KBSN that was proposed by [53]. The sub-

graph embedding module extracts the sub-graph closely connected to these core entities, while the sub-graph extraction module extracts the important text and picture features and maps them to the knowledge base (DBPedia). Finally, the sub-graph embedding module converts these sub-graphs into low-dimensional vectors. KAN [19] uses a co-attention mechanism. The model contains three modules: Top Edge Attention Module (Top-EAM), Central Attention Module (CAM), and bottom attention Module (Bottom – EAM). Glove embedded question features are forwarded through the CAM to obtain the attended question features. The attended visual feature is then created by passing the attended question features and visual features from the F-RCNN through the top EAM. The model uses ConceptNet to extract external knowledge about the important things in the image, and the bottom EAM gets the attended knowledge feature. The adaptive score module receives the attended image feature and the attended knowledge feature before sending its output to the classifier.Our study shows that (Figure 3) ConceptNet is the most commonly used knowledgebase as it is structured, relatively simple, smaller in size, and good for day-to-day common sense knowledge. Table 3 shows a comparison of various knowledge bases.

Table 3: Comparison of knowledgebases.

Knowledgebase	Advantages`	Limitations
ConceptNet	Structured, Relatively simple, smaller in size, Good for common sense day-to-day life reasoning.	Although covers most of the concepts related to general knowledge, depth and breadth of the coverage is less compared to other KBs.
Wikidata	Good for knowledge-driven open-ended task.	Unstructured and contains more noisy and redundant content due to its open-source nature, almost 10 times larger than ConceptNet
DBPedia	The DBpedia collects structured data from Wikipedia that spans a wide range of specialized fields and general knowledge.	Structured, but the knowledge that has been extracted is mainly restricted to named entities or concepts that have proper names, such as cities, people, species, films, organizations, etc. In DBpedia, there is no mention of the language relationship between these ideas, which is more important for ontology mappings.
WordNet	Wordnet can identify one or more Focuses on formal word taxonomies. In WordNet, one or more words may refer to synsets and synsets.	Focuses only on word taxonomies that are formal.

Source: Own elaboration.

Table 4: Knowledge bases used in literature.

KB	Method
WikiData	KVQA [49], Strings-To-Things [51], KAT [90]
DBPedia (st)	FVQA [45], KRISP [50], explainable VQR [52], KBSN [53]
ConceptNet (st)	KAN [19], Open domain VQ with DMN [38], ConceptBert [44], Augmented VQA [46], KINet [47], KRISP [50], explainable VQR [52], , MM answer validation [91]
WebChild	explainable VQR [52]
WordNet	KI-Net [47], TD VQA [63]
HasPart	KRISP [50]
Visual Genome	KINet [47]

Source: Own elaboration.

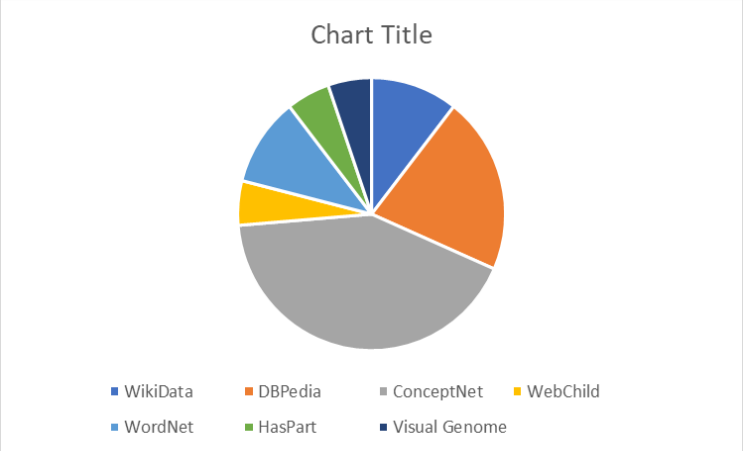

Figure 3: Distribution of usage of Knowledge Bases used in literature.
Source: Own elaboration.

VI. DOMAIN-SPECIFIC VQA

All of the VQA models that we discussed in the previous sections work on generic datasets that contain images from all domains. Designing a VQA model for a specific domain is still a challenging task due to the unavailability of the domain-specific dataset and domain-specific experts. Also, every domain requires different attributes in a dataset for example dataset in an educational image may require the inclusion of OCR tokens in a dataset along with image, question, and answer triplet. This section discusses how VQA has been applied to a specific domain and what are the challenges.

[54] proposed AI-based VQA system for pre-schoolchildren where the robot captures the image and identifies objects and automatically generates the question and answers. However, the authors kept the questions that are limited to colors and counting. [55] developed a VQA model for answering questions from charts like bar chart, line chart, pie chart etc. The authors used two datasets FigureQA [56] and DVQA [57]. To address inquiries concerning data visualization, they put forth a revolutionary algorithm. The system learns both the low-level and high-level properties of the image. It uses both low-level and high-level Q+I fusion. Also, all words in the image are extracted using OCR. Also introduced table reconstructions from charts by asking questions. Gracia et al. [58] offered a brand-new dataset named AQUA for analyzing artworks' answers to queries. Using artworks and comments included in the paintings, the question-answer (QA) pairings are created automatically. The model predicts the solution using the Painting and a paragraph that it has received from a knowledge repository. The model doesn't take into account variations in artistic styles. The visual recognition component would benefit from style correction tools. Additionally, extracting knowledge from numerous KBs may enhance efficiency. Many literature applied VQA on remote sensing images [59, 93]. [59] produced the 772 Images and 77,000 Questions and Answers in the new dataset RSVQA. OpenStreetMap (OSM) is used to gather the data needed to create the questions and answers. The model considered three types of tasks – classification, detection, and regression. For image features they used CNN and for question features they used RNN. A simple dot product is used for fusion. The challenge here is to obtain remote-sensing images with more resolution. And also improve the model with attention models. [60] SlideImage dataset that is being proposed for instructional usage. Data was gathered from the AI2D dataset and Wikimedia Commons, and test data was gathered from instructional slides. provided a dataset for image classification instruction. The VQA problem can be solved by expanding this dataset.

Another area where VQA is required is with regard to medical images. A few of the difficulties that VQA-Med encounters include the requirement for special processing of medical-specific vocabulary in medical texts and images, a challenge in combining multi-modal features at various levels of medical texts and images, and a propensity to ignore the relationship between the question and the visual information deduced from the text semantics. The VQA model was presented by [61] with two branches. The model uses a transformer structure for the common classification problem, three embedding methods, a hierarchy of feature extractors, a parallel structure of GRU and ResNet152 as image feature extractors, and specialized segmentation symbols as input. For irregular, open-ended questions without any workable candidate answers, this technique uses image retrieval to offer the text description answer that most closely fits the test image.

VII. VQA DATASETS

A number of datasets are available to experiment with visual Question answering challenges. This section discusses some of the important VQA datasets along with their merits and demerits. Nearly all of the VQA datasets contain a triplet containing a question, an image, and corresponding answers. Some of the datasets provide additional information such as Image captions, object-related facts, OCR tokens for the text present in an image, Bounding boxes for an image's objects etc. Most of the publicly available datasets have been created using a crowdsourcing mechanism [1,58,62] that used Amazon Mechanical Turk to collect questions about images and also respective answers. In [45] authors have appointed 38 human volunteers to create the dataset.

A. Generic VQA Datasets

COCO-QA

The COCO-QA dataset [63] is developed from Microsoft's COCO images. The number of images in the datasets is 1,23,287. There are 1,17,684 total questions that are created automatically from the image captions present in the COCO dataset, consisting of 78,786 training questions and 38,948 testing questions. The problem with this dataset is, it contains only four types of questions: object, counting, color, and location. Also, since the questions are automatically generated, they are not grammatically correct and all the answers are one–word.

DAQUAR

The earliest dataset for VQA is DAQUAR [42], which contains 1,449 images with 6,797 and 5,674 training and test question-answer pairs respectively. Questions and solutions were generated using both automated and manual annotations. The NYU depth V2 dataset served as the source of the images. The DAQUAR dataset was made available by the authors in two different forms: Full DAQUAR and Reduced DAQUAR. The Reduced version only comprises images from 37 classes with 3,825 training QA pairs and 297 testing QA pairs, while the Full dataset contains images from all 894 classes. There are only 25 test photos in it.

The major flaw of this dataset is its smaller size which is not adequate to train complex deep learning models. Also, the dataset contains most of the indoor scenes with poor lighting which makes it difficult to answer the question and even human accuracy on this dataset is approximately 50%.

VQA 1.0

This is the largest dataset available for VQA [1]. It contains both real and abstract images. There are 204721 images in all, 123287 of which were used for training and validation, and the remaining 81434 were retained for testing. All the images were taken from the Microsoft COCO dataset. It also contains 50000 abstract images. For abstract images, a train, Val, and test split is 20k, 10K, and 20K respectively. All the questions and annotations were human-generated. Three questions and ten human-generated answers are provided for each of the images. Multiple choice and open responses are the two types of answers offered by VQA V1. In open-answer mode, the model must choose the answer with the highest probability from all k potential responses, whereas in multiple-choice mode, the model must select the answer with the highest probability from the supplied choices (answers). With regard to the dataset, there are a few issues. Language bias affects the dataset. Most questions can be addressed only by asking more questions. Many questions lack a clear-cut answer because they are too subjective. Additionally, there aren't any questions in the dataset that need either strong reasoning or common sense.

VQA 2.0

The largest VQA dataset to date is this one. To make VQAv1 more balanced, [64] authors added more images. They included a complementary image for each image in VQAv1 so that the query Q still makes sense, but the answer is different. For instance, they added the complimentary picture I', for which the query Q still makes sense but the answer is A', to each triple of (I,Q,A):(image, Question, Answer). In essence, this gets rid of the language bias that was present in the original VQA sample. In total, there are 443K train, 214K validation, and 453K test (question, picture) pairings in the balanced VQA dataset, totaling 1.1M (image, question) pairs with roughly 13M related responses and 200K images.

CLEVR

CLEVR [65] is a dataset containing synthetic images of 3D objects of different shapes. Its training set contains 70,000 images and 699,989 questions. A validation set contains 15,000 images and 149,991 question-answer pair and a test set contains 15,000 images and 14,988 question-answer pair. This dataset was developed for testing the model's complex visual reasoning capabilities as questions in a dataset require a high level of reasoning. Although it is good for performing high-level complex reasoning, it is not suitable for real images as real images are natural and noisy.

Vizwiz

This dataset Vizwiz [21] was basically introduced to motivate the community to develop VQA systems for blind people to help them in their day-to-day life. Images in this dataset were taken by actual blind people via a Mobile app. It contains 32842 images and one question for each image. Answers are human-generated. The problem with this dataset is that images are of poor quality. and also, some questions are unanswerable due to poor quality.

SHAPES

SHAPES dataset [66] contains all the synthetic images of different colorful 3D shapes with 15,616 images and 244 questions with three types: attribute, relationship, and position. The dataset is good for testing mode's reasoning capability but it contains only yes/no kind of questions and does not generalize well for real-world images.

Visula7w

This dataset is made up of 47,300 images [67] picked up from the COCO dataset. 327,929 QA pairs, 1,311,756 multiple-choice questions created by humans, and 561,459 object groundings from 36,579 categories are all included in this collection. What, when, where, why, who, how, and which are the 7W questions in this dataset. It only provides multiple-choice answers.

B. Domain-Specific VQA datasets

Domain-specific datasets focus only on images of specific domains, for example, medical or education domains. These domain-specific VQA models have their own challenges. The following section discusses some of the popular domain-specific VQA datasets.

VQA – Med 2018: The VQA-Med dataset [68] was specially constructed for the medical field by using radiology images from the MedPix database. VQA-Med-2018 was the first-ever benchmark proposed in the field of medicine. It included a training set of 4,500 images and 4,500 question-answer (QA) pairs, a new validation set of 500 images and 500 QA pairs, and a new test set of 500 images and 500 questions about abnormality. Modality, plane, organ system, and abnormality were added as new question categories in later modified datasets called VQA-Med [2019] and VQA-Med [2020]. While the three classes—plane, modality, and organ system—can be solved as a classification task, the design of solutions for the fourth category, abnormality, offers a challenge. The dataset's images are rather noisy. Many questions and examples are also unhelpful for the management of patients. It is therefore irrational to include them in the dataset.

VQA - RAD: A radiology-specific dataset called VQA-RAD was proposed in 2018 [69]. The balanced image collection comes from the MedPix5 database and includes examples of the head, chest, and abdomen. The author showed the photographs to physicians to elicit open-ended queries in order to study the question in a realistic setting. Both free-from and template-structured inquiries must be created by the physicians. The QA pairs are then manually checked and categorized to examine the clinical focus. There are two different types of answers: closed-domain and open-domain. The VQA-RAD dataset, despite its small size, has gathered crucial knowledge regarding the questions an AI radiologist in a medical setting should be able to answer. The dataset is very small compared to other datasets and contains only 315 images with 3515 questions that makes use of deep learning techniques effective on this dataset.

FigureQA

FigureqQA dataset [56] is basically created for answering question from data visualization. Around 1,00000 images of variety of plots like bar charts, line graphs, and point-plot were generated synthetically. There are approximately one lac images and 1.3 million questions in a training set; the 20, 000 images with over 250, 000 questions in a validation and test set each. Images lack a lot of diversity present in real-world data visualizations because they were produced synthetically. The degree of diversity introduced is restricted by the features of the software program that created these photos. Additionally, the human-generated questions are not diverse enough.

DVQA

Like FigureQA, the DVQA [57] is also a data visualization dataset, but it contains only figures related to bar charts. The charts were made with Matplotlib. Structure, data, and reasoning questions are the three categories of questions included in DVQA, but there are just a handful of templates overall. It contains a total of 3487194 questions and 1576 unique answers, total 300000 of images.

AQUA

AQUA [58] is basically a dataset containing Art paintings and allows you to ask queries on these paintings. The question-answer pair is generated automatically using the question-generation methods from visual content of the painting and comments provided on the paintings. It contains both visual questions and knowledge-based questions. Apart from the visual image, and question, it also provides comments written on the painting to generate an answer. It contains 69,812 QA pairs among which 29568 are visual-based and 40244 are knowledge-based. Total of 19189 images with the split of 17117 training, 1032 testing, and 1040 for validation.

RSVQA

RSVQA [59] is basically a question-answering dataset that allows you to interact with Remote Sensing Images. The authors provided two datasets: LR- low resolution and HR-high resolution. It contains 10659 images and 955664 questions.

TEXT-VQA

Text-VQA [43] is the dataset that leverages text present in the images along with the question to produce the answer. Many times, the answer to the question may lie in the text present in the image. The dataset contains 28,408 images, with 45,336 questions and 453,360 ground truth answers. The questions in a dataset are such that the model needs to reason about the text present in the image to answer the question. The images were taken from OpenImages v3 dataset. The challenge here is to design an OCR model to extract text in the image since sometimes the text may be rotated or cropped or blurred.

C. Knowledge-Based VQA dataset

All of the above data sets discussed contain closed-domain questions, meaning the answer can be generated by processing the question and image itself. The open-domain questions are those which require the integration of external knowledge apart from the features given in the image. Many authors have attempted to create such challenging VQA datasets that require external and commonsense knowledge.

OKVQA

OKVQA [62] dataset contains questions that require external knowledge. Images were taken from COCO dataset and the queries and annotations were generated using human annotators. It contains 14031 images and 14,055 open-ended questions with 5 ground truth answers for each question. The questions are formed such that every question requires external knowledge to answer. Although small in size compared to VQA2.0, the questions in this dataset are more complex and challenging.

KVQA

Th2 KVQA dataset [49] contains the question-answers about the named entities such as Sachin Tendulkar or Barack Obama. It contains 24,602 images containing such name entities taken from Wikipedia pages, 183,007 QA pairs with 18,880 unique entities. The questions in this dataset require extracting knowledge about the multiple named entities present in the image and also the relationship between these entities. But this dataset contains only named entities and cannot be generalized for the real-world VQA.

FVQA

By adding supporting factual information to VQA samples, FVQA (Fact-based VQA) [70] essentially expands the current VQA collection. So now, in order to train the VQA model, four inputs are used: an image, a question, an answer, and a supporting fact. The supporting information is displayed as a triplet. The supporting fact is represented in the form of a triplet. 2,190 images were taken from the MSCOCO and contain 5826 questions. Using the tools and classifiers, each image is annotated with visual concepts (objects, scene, and actions). Each visual concept's knowledge is extracted from structured knowledge repositories like ConceptNet, DBpedia, and WebChild. Annotators created 5,826 questions that rely on the image for information and a few carefully selected facts to justify their answers. It requires a long training time and a large number of supporting facts also need to be trained. An efficient method of finding relevant facts from the huge factual database is required.

KB VQA

KB-VQA dataset [71] was created from 700 images chosen from the validation set of VQA2.0 such that overall 150 different object classes and around 100 scene classes were covered. The dataset contains three categories of questions:1256 visual questions, 883 common sense questions, and 263 KB-knowledge questions. So, 505 of the questions require external knowledge. The problem here is dataset is relatively small compared to the original VQA dataset.

A-OKVQA

A-OKVQA [22] is a successor of OK-VQA. The dataset includes questions that call for the use of a range of knowledge forms, including common sense, general knowledge, and visual knowledge. The dataset contains 23692 images taken from COCO image dataset with 24,903 question, answer, and fact triples.

Table 5: Comparison of various datasets available for VQA

Dataset	Source	No. of Images	No. of questions	Annotations (Humans/Auto)	Number of question type/categories	Answer Type	Average Answer length	Answers	Evaluation Metric	Limitations
COCO-QA	MSCOCO	123287	117684	Automatic generated from image caption	4 (Object, Number, Color, Location)	Open-ended	1.1	One word	Accuracy, WUPS	Not proper phrasing of questions, Grammatical errors in questions, Questions are not equally distributed among the 4 types.
DAQUAR	NYU- Depth v2	1449	Q- 12468	Human + Automatic	4	open-ended	1.1	One word	Accuracy, WUPS, Consensus	Contains indoor scenes that make answering difficult. Also too small in size.
VQA- Real	MSCOCO	2,04,721	Q-614613, 7,984,119	Human	20	Open, MCQ	1.2	One , two, or three word	Accuracy, Consensus	Suffers from language Bias. Most of the questions can be answered using only questions. Many questions are too subjective to have a single answer.
VQA – Balanced	Clipatr	15,623	33379	Human	20+	Open, MCQ	1.2	One , two, or three word	Accuracy	Although better balanced than VQA1.0, still the dataset is skewed with respect to distribution of the type of questions.
CLEVR	Rendered using Blender	100000	864968	Automatic	90 Question family	open-ended	1	One word	Accuracy	Since images are synthetic, cannot be generalized to a real word setting.
Vizwiz	Consists of images taken by blind	32842	32842	Human	--	Open-ended	1	Multiple words	Accuracy, Consensus	Images are of poor quality. Also, some of the questions are unanswerable.
SHAPES	Synthetic images	15,616	244	Human	3 (Attribute, Relationship, Position)	Yes/No	1	One word (Yes/No)	Accuracy	Only yes/no questions, Cannot be used to generalize real world images
Visual7W	Subset of Visual Gnome	47300	47,300	Human	7 W questions	MCQ	2.2	Multi choice answer mode	Accuracy	Provides only Multiple choice answer Mode.

Source: Own elaboration

Aibi research, management and engineering journal. Volume 11, Number 3, September - December 2023, ISSN 2346-030X PP 126-142.

Dataset	Source	No. of Images	No. of questions	Annotations (Humans/ Auto)	Number of question type/categories	Answer Type	Average Question length	Average Answer length	Answers	Evaluation Metric	Limitations
VQA-Med	MedPix	5000	5000	semi-automatic	11/ (4)	Open – ended	—	—	Multi word	Accuracy, BLEU, METEOR	Images are highly noisy. Some of the images are questions are irrelevant to the patient's treatment. Some of the questions are unanswerable
VQA-RAD	Contains radiology images taken	315	3515	Questions are generated form Clinicians	11	Open-ended, MCQ	5 to 7 words	—	Multi word	Accuracy, BLEU	Relatively small dataset.
Text-VQA	OpenImages v3 dataset	28,408	45,336	Human	Text based	Open ended	7.1667	1.7	Multi word (paragraph or quotes)	Accuracy	Needs better OCR techniques to get good results
FigureQA	synthetically generated with Bokeh	120000	1.4Million	Auto	15	Yes/No	—	1 word	single word yes/No	Human Judgement	provides only Yes/No type of questions, lacks variability generated from real-world data visualization, lacks diversity of human generated complex queries.
DVQA	Synthetically generated using matplotlib	300,000 images	3,487,194 total question-answer pairs	Auto	Three: structure understanding, data retrieval and reasoning	—	—	one word	Mostly one word	% of questions answered correctly.	Includes only bar charts but provides variations in question compared to FigureQA. lacks variability generated from real-world data visualization
AQUA	SemART dataset	19189	69812 Generated using question -answer generation method	auto later refined by Humans	two: visual and knowledge based	Open	8.82	3.13	visual: 1 word, Knowledge-Based: 3 to 4 words	EM - exact match	Since QA pairs are automatically generated, it lacks human generated complex questions.
RSVQA	10659	USG's High resolution ortho rectified images	9, 55664QA derived from OSM	automatic	three: account, presence, comparison, rural/urban	Open	—	—	—	accuracy	Some annotations in datasets are missing or badly registered. since QA are generated automatically they are limited and not as complex as real - world human generated.

Dataset	Source	No. of Images	No. of questions	Annotations (Humans/Auto)	Number of question type/categories	Answer Type	Average Question length	Average Answer length	Evaluation Metric	Evaluation Metric	Limitations
FVQA	MSCOCO	2190	5826	Human	32	Open ended	9.5	1.2	Accuracy, WUPS,Human Judgment	Accuracy, WUPS,Human Judgment	Long Training time
KB-VQA	MSCOCO +Imagenet	700	2402	Human	3 types - Common sense, Visual and KB	Open ended	6.8	2	Human Judgment	Human Judgment	Relatively Small dataset
KVQA	WikiPedia Pages, Wikidata	24k	130K	Human	–	Open ended	10.14	1.64	Accuracy	Accuracy	Works only for named entities
OKVQA	MSCOCO	14031	14055	Human	10 - Vehicles and Transportation; Brands, Companies and Products;	Open ended	8.1	1.3	accuracy	accuracy	An improvement over all other VQA datasets in terms of question quality
A-OKVQA	MSCOCO	23692	24903	Human	4 - common sense, Physical, Knoowledge base, Visual	open - ended, MCQ	8.8	1.3	accuracy	accuracy	__

Aibi research, management and engineering journal. Volume 11, Number 3, September - December 2023, ISSN 2346-030X PP 126-142.

VIII. EVALUATION METRIC

Various evaluation measures have been proposed in the literature to evaluate VQA models. The most commonly used metric is VQA accuracy. Other useful metrics are BLUE, WUPS, METEOR, and Human judgment.

Accuracy:

Antol S. [1] has proposed a new accuracy measure to evaluate the VQA model that takes care of variation in human answers. This is also called accuracy based on human consensus. The formula for computing accuracy is,

Accuracy = min (number of people agreed upon the model generated answer/3,1)

This means that the output is considered accurate if at least three human agreements are there.

WUPS (Wu and Palmer Similarity)

Wu et al. [72] suggested an alternative metric called WUPS that can be used to measure the accuracy of the VQA model. It makes use of semantic similarity between the generated answer and the actual answer to check how far the predicted answer is from the correct answer. WUPS takes the model-generated answer and ground truth answer as inputs and generates the score between 0 to 1 based on the List Common subsumer between the two. It assumes that the path length and depth in path-based measures determine how related two concepts are to one another. So, words that are totally different but have the same semantic meaning will be less penalized here. For example, the word "chair" and "wooden chair" will have a similarity, of 0.96 but "chair" and "furniture" will have 80% similarity. But, WUPS has a problem that, it may assign a high value to distant concepts. For example, "sea" and "water" may be assigned 40% similarity. A Solution to this is to use a threshold WUPS score, where a score below a certain threshold is scaled down by certain factors. A threshold of 0.9 and a scale of 0.1 was suggested by [94]. Many authors [3,49,54] have used this threshold WUPS in addition to normal accuracy to measure model performance. Another problem with WUPS is that it generates high scores for similar concepts, this creates a problem measuring the answer related to the attribute of an object. For example, answers "Red" and "Yellow" will have high scores as they are semantically close to each other.

BLEU (Bilingual Evaluation Understudy)

BLEU [73] is another significant statistic that many authors [81,82] utilized. It also produces output between 0 and 1, just like WUPS. With values closer to 1, there is a greater similarity between the expected and actual results, which is indicated by this number. The approach determines the proportion of matching n-grams between the ground truth answer and the anticipated response. Regardless of word order, the comparison is made.

$$P_n = \frac{\sum_{n-grams} count_{clip}\,(n\ -\ gram)}{\sum_{n-grams} count(n-gram)}$$

METEOR

The METEOR [74] (Metric for evaluation of translation with explicit ordering) metric takes into account multiple aspects of translation quality, including precision, recall, stemming, and synonymy. It is designed to address some of the limitations of the BLEU metric, such as its reliance on n-gram precision and insensitivity to word order. To calculate the METEOR, first, calculate the precision and recall for all unigrams. Then it computes a harmonic mean of the precision and 9 times the recall.

$$M_{mean} = \frac{10P}{P + 9R}$$

The following is how METEOR calculates a penalty for a particular alignment in order to account for longer matches. First, all the unigrams in the system translation that corresponds to unigrams in the reference translation are grouped into the fewest number of chunks possible, ensuring that each chunk's unigrams are in adjacent positions in both the system translation and the reference translation, where they are also mapped to unigrams in adjacent positions.

The second part is a penalty function that is formulated as follows:

$$Penalty\ =\ 0.5\ x\ \left(\frac{no.of\ chunks}{no.of\ unigrams}\right)^3$$

Finally, the score is computed as,

$$s\ =\ M_{mean}\ x\ (1-\ penalty)$$

Human Judgement:

The most reliable and accurate method of evaluation is human judgment. All the model-generated answers are given to humans for evaluation. But this method is quite expensive and time-consuming. This can be used as a secondary method of evaluation.

IX. DISCUSSION AND CONCLUSION

This survey presents a comprehensive review of four types of VQA models: Simple joint embedding-based VQA models, Attention-based VQA models, Knowledge Incorporated VQA models, and domain-specific VQA models. In detail, three categories of datasets: generics datasets, Domain-specific datasets, and Knowledge incorporated datasets have been discussed along with their merits and demerits. Finally, some of the important guidelines deduced from the survey have been presented.

- Most of the VQA models discussed above are converted into classification models by considering a top 'n' number of frequently occurring answers. VQA is actually a generation problem rather than a classification. Very few pieces of literature treat VQA as a generation problem.
- Existing metrics are still not sufficient and accurate to evaluate VQA models due to the lack of considering human bias in answering a question, thus there is a need to develop new more accurate metrics for evaluating VQA models.
- Most of the VQA models fail where there is a question based on complex reasoning and outside knowledge posed. Still, attempts have been made to integrate external knowledge from various sources. A robust method of integrating external knowledge and searching techniques may help improve the model's answer prediction capabilities.
- Local features provide richer information about objects than global features. Combining global and local features together may further improve model performance.
- Attention mechanism is more effective as compared to simple joint embedding techniques. Further improvement in the co-attention mechanism will allow more interaction between question and image features.
- All the latest literature use transformers for both vision and language modeling. Using transformers for both vision and language models allows to extract of rich contextual information from image and questions.
- Incorporating external knowledge into the visual question-answering model is still a challenge. The most commonly used KB is conceptNet due to its simplicity and ease of use. Moreover, combining knowledge from multiple external knowledge bases may help model extracting a variety of concepts.
- There is still a challenge working on domain-specific datasets. The creation of a vast domain-specific dataset and developing a suitable model that works for that dataset is an open challenge.

X. REFERENCES

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, 2015. VQA: visual question answering, Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2425–2433. doi: 10.1109/ICCV.2015.279.

[2] Nelson Ruwa, Qirong Mao, Liangjun Wang, Ming Dong, 2018, Affective Visual Question Answering Network, IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 2018, pp. 170-173, doi: 10.1109/MIPR.2018.00038.

[3] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. 2015. Ask Your Neurons: A Neural-Based Approach to Answering Questions about Images. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV) (ICCV '15). IEEE Computer Society, USA, 1–9. https://doi.org/10.1109/ICCV.2015.9.

[4] Geonmo Gu, Seong Tae Kim, Yong Man Ro, 2017, Adaptive attention fusion network for visual question answering, IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 2017, pp. 997-1002, doi: 10.1109/ICME.2017.8019540.

[5] Ilija Ilievski, Shuicheng Yan, Jiashi Feng,206. A Focused Dynamic Attention model for visual question answering' [Online]. Available: https://arxiv.org/abs/-1604.01485.

[6] Zichao Yang; Xiaodong He; Jianfeng Gao; Li Deng; Alex Smola, 2016. Stacked Attention Networks for Image Question Answering, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 21-29.

[7] Kushal Kafle; Christopher Kanan, 2016. Answer-Type Prediction for Visual Question Answering, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4976-4984.

[8] Duy-Kien Nguyen; Takayuki Okatani,2018. Improved Fusion of Visual and Language Representations by Dense Symmetric Co-attention for Visual Question Answering," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 6087-6096, doi: 10.1109/CVPR.2018.00637.

[9] Deepak Gupta, Pabitra Lenka, Asif Ekbal, Pushpak Bhattacharyya, 2020. A Unified Framework for Multilingual and Code-Mixed Visual Question Answering, Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing (pp. 900–913). Association for Computational Linguistics.

[10] Jiasen Lu, Jianwei Yang, Dhruv Batra, Devi Parikh, 2016. Hierarchical question image co-attention for visual question answering,' in Proc. NIPS, 2016, pp. 289_297.

[11] Chao Yang; Mengqi Jiang; Bin Jiang; Weixin Zhou; Keqin Li,2019. Co-Attention Network with Question Type for Visual Question Answering, IEEE Access, vol. 7, pp. 40771-40781, doi: 10.1109/ACCESS.2019.2908035.

[12] Lianli Gao, Liangfu Cao, Xing Xu, Jie Shao, Jingkuan Song,2020. Question-Led object attention for visual question answering, ,Neurocomputing, Volume 391, 2020, Pages 227-33, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2018.11.102.

[13] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, Article 2, 13–23.

[14] Hao Tan , Mohit Bansal, 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 5100–5111, Hong Kong, China, November 3–7, 2019.

[15] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, Jingjing Li u, 2020. UNITER: UNiversal Image-TExt Representation Learning, https://doi.org/10.48550/arXiv.1909.11740.

[16] Xiujun Li, Xi Yin, Chunyuan Li , Pengchuan Zhang , Xiaowei Hu , Lei Zhang, Lijuan Wang , Houdong Hu , Li Dong , Furu Wei , Yejin Choi , and Jianfeng Gao ,,2020. OSCAR : Object-Semantics Aligned Pre-training for Vision-Language Task, Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science, vol 12375. Springer, Cham. https://doi.org/10.1007/978-3-030-58577-8.

[17] Ze Hu; Jielong Wei; Qingbao Huang; Hanyu Liang; Xingmao Zhang; Qingguang Liu , 2020. Graph Convolutional Network for Visual Question Answering Based on Fine-grained Question Representation," 2020 IEEE Fifth International Conference on Data Science in Cyberspace (DSC), Hong Kong, China, 2020, pp. 218-224, doi: 10.1109/DSC50466.2020.00040.

[18] Kevin J. Shih; Saurabh Singh; Derek Hoiem, 2016,. Where to Look: Focus Regions for Visual Question Answering, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4613-4621, doi: 10.1109/CVPR.2016.499.

[19] Liyang Zhang , Shuaicheng Liu , Donghao Liu, Pengpeng Zeng, Xiangpeng Li, Jingkuan Song , Lianli Gao, 2021. Rich Visual Knowledge-Based Augmentation Network for Visual Question Answering, in IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 10, pp. 4362-4373, Oct. 2021, doi: 10.1109/TNNLS.2020.3017530.

[20] Peng Zhang; Yash Goyal; Douglas Summers-Stay; Dhruv Batra; Devi Parikh,2016, Yin and Yang: Balancing and Answering Binary Visual Questions, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 5014-5022, doi: 10.1109/CVPR.2016.542.
[21] Danna Gurari; Qing Li; Abigale J. Stangl; Anhong Guo; Chi Lin; Kristen Grauman; Jiebo Luo; Jeffrey P. Bigham, 2018. VizWiz Grand Challenge: Answering Visual Questions from Blind People, IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 3608-3617, doi: 10.1109/CVPR.2018.00380.
[22] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, Roozbeh Mottaghi 2022, A-OKVQA: A Benchmark for Visual Question Answering using World Knowledge, In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds) Computer Vision – ECCV 2022. ECCV 2022. Lecture Notes in Computer Science, vol 13668. Springer, Cham. https://doi.org/10.1007/978-3-031-20074-8_9.
[23] Binh X. Nguyen; Tuong Do; Huy Tran; Erman Tjiputra; Quang D. Tran; Anh Nguyen,2022, Coarse-to-Fine Reasoning for Visual Question Answering, IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, 2022, pp. 4557-4565, doi: 10.1109/CVPRW56347.2022.00502.
[24] Caiming Xiong, Stephen Merity, Richard Socher ,2016. DMN: Dynamic Memory Networks for Visual and Textual Question Answering, https://arxiv.org/abs/1603.01417v1.
[25] M. Dias, H. Aloj, N. Ninan and D. Koshti, "BERT based Multiple Parallel Co-attention Model for Visual Question Answering," 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2022, pp. 1531-1537, doi: 10.1109/ICICCS53718.2022.9788253.
[26] Dipali Koshti, Ashutosh Gupta, and Mukesh Kalla, 2022, BERT based Hierarchical Alternating Co-Attention Visual Question Answering using Bottom-Up Features", Int J Intell Syst Appl Eng, vol. 10, no. 3s, pp. 158–168, Dec. 2022. https://doi.org/10.17762/ijisae.v10i3S.2427.
[27] Bei Liu, Zhicheng Huang, Zhaoyang Zeng, Zheyu Chen, Jianlong Fu, 2019. Learning Rich Image Region Representation for Visual Question Answering., Learning Rich Image Region Representation for Visual Question Answering, ArXiv, abs/1910.13077.
[28] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang,2018. Bilinear attention networks. In Proceedings of the 32nd International Conference on Neural Information Processing Systems (pp. 1571–1581). Curran Associates.
[29] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach, 2017. Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847, 2016
[30] Jin-Hwa Kim, Sang-Woo Lee, Dong-Hyun Kwak, Min-Oh Heo, Jeonghee Kim, JungWoo Ha, and Byoung-Tak Zhang. Multimodal residual learning for visual QA. In Proceedings of the 30th International Conference on Neural Information Processing Systems (pp. 361–369). Curran Associates Inc.
[31] Alberto Mario Bellini, 2020. Towards Open-Ended VQA Models Using Transformers, University of Illinois at Chicago https://doi.org/10.25417/UIC.13475892.V1.
[32] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao, 2017. Multimodal factorized bilinear pooling with co-attention learning for visual question answering. In ICCV, pages 1839–1848, DOI:10.1109/ICCV.2017.202.
[33] Junwei Liang, Lu Jiang, Liangliang Cao, Li-Jia Li, and Alexander G. Hauptmann, 2018. Focal visual-text attention for visual question answering. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 6135-6143, doi: 10.1109/CVPR.2018.00642.
[34] Ahmed Osman, Wojciech Samek, DRAU: Dual Recurrent Attention Units for Visual Question Answering, Computer Vision and Image Understanding, Volume 185, 2019,Pages 24-30,ISSN 1077-3142, https://doi.org/10.1016/j.cviu.2019.05.001.
[35] Sheng Zhang, Min Chen, Jincai Chen, Fuhao Zou, Yuan-Fang Li, Ping Lu, 2021.Multimodal feature-wise co-attention method for visual question answering, Information Fusion,Volume 73,2021,Pages 1-10,ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2021.02.022.
[36] Hedi Ben-younes, Remi Cadene, Matthieu Cord, Nicolas Thome, 2017. MUTAN:Multimodal Tucker Fusion for Visual Question Answering, 2631-2639. 10.1109/ICCV.2017.285.
[37] Farazi Moshiur, Salman Khan, Nick Barnes, 2020. Attention Guided Semantic Relationship Parsing for Visual Question Answering. https://doi.org/10.48550/arXiv.2010.01725.
[38] Guohao Li, Hang Su, Wenwu Zhu, 2017. Incorporating External Knowledge to Answer Open-Domain Visual Questions with Dynamic Memory Networks, ArXiv abs/1712.00733. https://doi.org/10.48550/arXiv.1712.00733.
[39] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, Weidi Xie, "2023, PMC-VQA: Visual Instruction Tuning for Medical Visual Question Answering: https://arxiv.org/abs/2305.10415v5.
[40] Fuji Ren; Yangyang Zhou,2020, CGMVQA: A New Classification and Generative Model for Medical Visual Question Answering," in IEEE Access, vol. 8, pp. 50626-50636, 2020, doi: 10.1109/ACCESS.2020.2980024.
[41] Weixin Liang, Yanhao Jiang, Zixuan Liu, GraghVQA: Language-Guided Graph Neural Networks for Graph-based Visual Question Answering, https://doi.org/10.48550/arXiv.2104.10283.
[42] Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Mingkui Tan, Chuang Gan, 2020, Location-Aware Graph Convolutional Networks for Video Question Answering, Proceedings of the AAAI Conference on Artificial Intelligence, 34(07), 11021-11028. https://doi.org/10.1609/aaai.v34i07.6737.
[43] Damien Teney, Lingqiao Liu, Anton van den Hengel, Graph-Structured Representations for Visual Question Answering," 2017. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 3233-3241, doi: 10.1109/CVPR.2017.344.
[44] François Gardères, Maryam Ziaeefard, Baptiste Abeloos, Freddy Lecue2020., "ConceptBert: Concept-Aware Representation for Visual Question Answering, In Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 489–498). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2020.findings-emnlp.44.
[45] Peng Wang, Qi Wu,Chunhua Shen, Anthony Dick, Anton van den Hengel, 2017. FVQA: Fact-based Visual Question Answering" , in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 10, pp. 2413-2427, 1 Oct. 2018, doi: 10.1109/TPAMI.2017.2754246.
[46] Maryam Ziaeefard and Freddy Lecue,2021, Towards Knowledge-Augmented Visual Question Answering, In Proceedings of the 28th International Conference on Computational Linguistics (pp. 1863–1873). International Committee on Computational Linguistics.
[47] Yifeng Zhang, Ming Jiang, Qi Zhao, 2021. Explicit Knowledge Incorporation for Visual Reasoning, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 1356-1365, doi: 10.1109/CVPR46437.2021.00141.

[48] Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, Anton van den Hengel,2018,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, 1 June 2018, doi: 10.1109/TPAMI.2017.2708709.

[49] Sanket Shah , Anand Mishra , Naganand Yadati , Partha Pratim Talukdar , 2019. KVQA: Knowledge-Aware Visual Question Answering. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 8876-8884. https://doi.org/10.1609/aaai.v33i01.33018876.

[50] Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta,Marcus Rohrbach, 2021.KRISP: Integrating Implicit and Symbolic Knowledge for Open-Domain Knowledge-Based VQA, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 14106-14116, doi: 10.1109/CVPR46437.2021.01389

[51] Ajeet Kumar Singh, Anand Mishra, Shashank Shekhar, Anirban Chakraborty,2019. From Strings to Things: Knowledge-enabled VQA Model that can Read and Reason, IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 4601-4611, doi: 10.1109/ICCV.2019.00470.

[52] Qingxing Cao, Bailin Li, Xiaodan Liang and Liang Lin,2019. Explainable High-order Visual Question Reasoning: A New Benchmark and Knowledge-routed Network, ArXiv, abs/1909.10128.

[53] Wenfeng Zhenga, Lirong Yinb, Xiaobing Chena, Zhiyang Maa ,Shan Liua , Bo Yanga, 2021, Knowledge base graph embedding module design for Visual question answering model", Pattern Recognition, Volume 120, 2021, 108153, ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2021.108153.

[54] Bin He,Meng Xia, Xinguo Yu, Pengpeng Jian, 2017. An educational robot system of visual question answering for preschoolers. 441-445. 10.1109/ICRAE.2017.8291426.

[55] Kushal Kafle, Robik Shrestha, Brian Price, Scott Cohen, Christopher Kanan, 2020. Answering Questions about Data Visualizations using Efficient Bimodal Fusion, IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass, CO, USA, 2020, pp. 1487-1496, doi: 10.1109/WACV45572.2020.9093494.

[56] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Akos Kadar, Adam Trischler, Yoshua Bengio , 2018, FigureQA: An Annotated Figure Dataset for Visual Reasoning, ICLR 2018.

[57] Kushal Kafle, Brian Price, Scott Cohen, Christopher Kanan,,2018. DVQA: Understanding Data Visualizations via Question Answering,IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 5648-5656, doi: 10.1109/CVPR.2018.00592.

[58] Noa Garcia, Chentao Ye, Zihua Liu, Qingtao Hu, Mayu Otani, Chenhui Chu, Yuta Nakashima, Teruko Mitamura, 2020. AQUA: A Dataset and Baselines for Visual Question Answering on Art, ECCV Workshop, 2020 Springer, https://arxiv.org/abs/2008.12520v1.

[59] Sylvain Lobry, Diego Marcos, Jesse Murray, Devis Tuia, 2020 ,RSVQA: Visual Question Answering from Remote Sensing Data, in IEEE Transactions on Geoscience and Remote Sensing, vol. 58, no. 12, pp. 8555-8566, Dec. 2020, doi: 10.1109/TGRS.2020.2988782.

[60] David Morris, Eric Budack,2020. SlideImages: A Dataset for Educational Image Classification, Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part IIApr 2020Pages 289–296 https://doi.org/10.1007/978-3-030-45442-5.

[61] Shengyan Liu, Xuejie Zhang, Xiaobing Zhou & Jian Yang , 2022,BPI-MVQA: a bi-branch model for medical visual question answering. BMC Med Imaging 22, 79 (2022). https://doi.org/10.1186/s12880-022-00800-x.

[62] Kenneth Marino; Mohammad Rastegari; Ali Farhadi; Roozbeh Mottaghi, 2019, OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 3190-3199, doi: 10.1109/CVPR.2019.00331.

[63] Mengye Ren, Ryan Kiros, and Richard S. Zemel, 2015. Exploring models and data for image question answering. In NIPS, 2015.

[64] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh, 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 6325-6334, doi: 10.1109/CVPR.2017.670.

[65] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick, 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 1988-1997, doi: 10.1109/CVPR.2017.215.

[66] Andreas, J., Rohrbch M., Darrell T., & Klein D. 2016. Deep compositional question answering with neural module networks, in: CVPR, 2016.In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 39–48).

[67] Yuke Zhu, Oliver Groth, Michael S. Bernstein, and Li FeiFei, 2016. Visual7w: Grounded question answering in images, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 4995-5004, doi: 10.1109/CVPR.2016.540.

[68] Asma Ben Abacha, Vivek V. Datla, Sadid A. Hasan, Dina Demner-Fushman, & Henning Muller, 2020. Overview of the VQA-Med Task at ImageCLEF 2020: Visual Question Answering and Generation in the Medical Domain. In CLEF 2020 Working Notes. CEUR-WS.org.

[69] Jason J. Lau, Soumya Gayen, Asma Ben Abacha Dina Demner-Fushman, 2018. A dataset of clinically generated visual questions and answers about radiology images. Sci Data 5, 180251 (2018). https://doi.org/10.1038/sdata.2018.251.

[70] Peng Wang, QiWu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. 2016, FVQA: Fact-Based Visual Question Answering," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 10, pp. 2413-2427, 1 Oct. 2018, doi: 10.1109/TPAMI.2017.2754246.

[71] Peng Wang, Qi Wu, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel, 2017. Explicit knowledge-based reasoning for visual question answering, Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence Main track. Pages 1290-1296. https://doi.org/10.24963/ijcai.2017/179.

[72] Zhibiao Wu and Martha Palmer,1994. Verbs semantics and lexical selection, in: Proceedings of the 32nd annual meeting on Association for Computational Linguistics, Association for Computational Linguistics, 1994, pp. 133–138

[73] Kishore Papineni, Salim Roukos, Todd Ward and WeiJing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02). Philadelphia, PA. July 2002. pp. 311-318.

[74] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 65–72, Ann Arbor, Michigan.

[75] Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han, 2016, Image Question Answering Using Convolutional Neural Network with Dynamic Parameter Prediction," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 30-38, doi: 10.1109/CVPR.2016.11.

[76] Kan Chen, JiagWang, Liang-Chieh Chen, Haoyuan Gao,Wei Xu, and Ram Nevatia.2015. ABC-CNN: An Attention Based Convolutional Neural Network for Visual Question Answering. arXiv preprint arXiv:1511.05960.
[77] Teney D, Hengel AV (2018) Visual question answering as a meta learning task. In: Computer vision—ECCV 2018 lecture notes in computer science. 229–245. Teney, D., & Hengel, A.V. (2017). Visual Question Answering as a Meta Learning Task. ArXiv, abs/1711.08105.
[78] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein, 2016. Neural Module Networks IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 39-48, doi: 10.1109/CVPR.2016.12.
[79] Bei Liu, Zhicheng Huang, Zhaoyang Zeng, Zheyu Chen, Jianlong Fu , 2019. Learning Rich Image Region Representation for Visual Question Answering, arXiv:1910.13077v1.
[80] Yang Shi, Tommaso Furlanello, Sheng Zha, Animashree Anandkumar ,2018. Question type guided attention in visual question answering. Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IVSep 2018Pages 158–175 https://doi.org/10.1007/978-3-030-01225-0_10.
[81] Duy-Kien Nguyen and Takayuki Okatani, 2018. Improved fusion of visual and language representations by dense symmetric co-ttention for visual question answering. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 6087-6096, doi: 10.1109/CVPR.2018.00637.
[82] Haoyuan gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu, 2015. Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering. In Proc. Advances in Neural Inf. Process. Syst. https://doi.org/10.48550/arXiv.1505.05612.
[83] Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus, 2015. Simple baseline for visual question answering. arXiv preprint arXiv:1512.02167.
[84] Huijuan Xu and Kate Saenko, 2015. Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering. arXiv preprint arXiv:1511.05234.
[85] Pan Lu, Hongsheng Li, Wei Zhang, Jianyong Wang, and Xiaogang Wang, 2018. Co-attending free-form regions and detections with multi-modal multiplicative feature embedding for visual question answering", In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence. AAAI Press.
[86] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, Anton van den Hengel, 2017. Explicit knowledge-based reasoning for visual question answering. In IJCAI, pages 1290–1296, https://doi.org/10.24963/ijcai.2017/179.
[87] Wenfeng Zheng, Lirong Yin, Xiaobing Chen, Zhiyang Ma, Shan Liu, Bo Yang, 2021.Knowledge base graph embedding module design for Visual question answering model, Pattern Recognition, Volume 120, 2021, 108153, ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2021.108153.
[88] Prashan Wanigasekara; Kechen Qin; Emre Barut; Fan Yang; Weitong Ruan; Chengwei Su, 2022. "Semantic VL-BERT Visual Grounding via Attribute Learning," 2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy, 2022, pp. 1-8, doi: 10.1109/IJCNN55064.2022.9892420.
[89] Damien Teney, Peter Anderson, Xiaodong He, Anton van den Hengel, 2017. Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge, IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 4223-4232, doi: 10.1109/CVPR.2018.00444.
[90] Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander Hauptmann, Yonatan Bisk, and Jianfeng Gao. KAT: A knowledge augmented transformer for vision-and-language. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, July 2022.
[91] Jiasen Lu, Ashish Sabharwal, Roozbeh Mottaghi,2022. Multi-Modal Answer Validation for Knowledge-Based VQA. Proceedings of the AAAI Conference on Artificial Intelligence, 36(3), 2712-2721. https://doi.org/10.1609/aaai.v36i3.20174.
[92] Hyeonwoo Noh, Taehoon Kim, Jonghwan Mun, Bohyung Han, 2019. Transfer Learning via Unsupervised Task Discovery for Visual Question Answering, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 8377-8386, doi: 10.1109/CVPR.2019.00858.
[93] Xiangtao Zheng; Binqiang Wang; Xingqian Du; Xiaoqiang Lu, Mutual Attention Inception Network for Remote Sensing Visual Question Answering, in IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-14, 2022, Art no. 5606514, doi: 10.1109/TGRS.2021.3079918.
[94] Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'14). MIT Press, Cambridge, MA, USA, 1682–1690.