IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 14 October 2023, accepted 4 November 2023, date of publication 20 November 2023,
date of current version 8 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3335216

SURVEY

A Critical Analysis of Benchmarks, Techniques, and Models in Medical Visual Question Answering

SUHEER AL-HADHRAMI[1,2], MOHAMED EL BACHIR MENAI[1], SAAD AL-AHMADI[1], AND AHMED ALNAFESSAH[3,4]

[1]Computer Science Department, King Saud University, Riyadh 12371, Saudi Arabia
[2]Department of Computer Engineering, Hadhramout University, Al Mukalla 10587, Yemen
[3]King Abdulaziz City for Science and Technology, Riyadh 11442, Saudi Arabia
[4]Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Corresponding author: Ahmed Alnafessah (aalnafessah@kacst.edu.sa)

This work was supported in part by the Research Center of the College of Computer and Information Sciences [King Saud University (KSU)], in part by the Center for Complex Engineering Systems [jointly between Massachusetts Institute of Technology (MIT) and King Abdulaziz City for Science and Technology (KACST)], and in part by the Deanship of Scientific Research at King Saud University for funding and supporting this Research through the Initiative of Deanship of Scientific Research (DSR) Graduate Students Research Support (GSR).

ABSTRACT This paper comprehensively reviews medical VQA models, structures, and datasets, focusing on combining vision and language. Over 75 models and their statistical and SWOT (Strengths, Weaknesses, Opportunities, Threats) analyses were compared and analyzed. The study highlights whether the researchers in the general field influence those in the medical field. According to an analysis of text encoding techniques, LSTM is the approach that is utilized the most (42%), followed by non-text methods (14%) and BiLSTM (12%), whereas VGGNet (40%) and ResNet (22%) are the most often used vision methods, followed by Ensemble approaches (16%). Regarding fusion techniques, 14% of the models employed non-specific methods, while SAN (13%) and concatenation (10%) were frequently used. The study identifies LSTM-VGGNet and LSTM-ResNet combinations as the primary approaches in medical VQA, with 18% and 15% usage rates, respectively. The statistical analysis of medical VQA from 2018 to 2023 and individual yearly analyses reveals consistent preferences for LSTM and VGGNet, except in 2018 when ResNet was more commonly used. The SWOT analysis provides insights into the strengths and weaknesses of medical VQA research, highlighting areas for future exploration. These areas include addressing limited dataset sizes, enhancing question diversity, mitigating unimodal bias, exploring multi-modal datasets, leveraging external knowledge, incorporating multiple images, ensuring practical medical application integrity, improving model interpretation, and refining evaluation methods. This paper's findings contribute to understanding medical VQA and offer valuable guidance for future researchers aiming to make advancements in this field.

INDEX TERMS Attention, deep learning, NLP, QA, question answering, SWOT analysis, vision, vision and language, visual, VQA.

I. INTRODUCTION

Question answering (QA) is a process used to answer questions written in a natural language. When these questions focus on visual information, the process is called visual question answering (VQA). The VQA about non-medical images is called general VQA, whereas the VQA about medical images is called medical VQA. This paper focuses on medical images and related VQA questions. VQA is a multidisciplinary task that involves natural language processing (NLP), computer vision (CV), and knowledge representation and reasoning (KR). The VQA chart in Figure 1 asks whether the image contains fundus exudates, extracts the features from the image (CV) and text (NLP), and interprets the relationship between them (KR) to answer the question by a classifier.

The associate editor coordinating the review of this manuscript and approving it for publication was Wenbing Zhao.

Recently, given the high significance of deep learning and the use of transfer learning in building vision and NLP models, VQA has become a challenge for artificial intelligence (AI) researchers. VQA helps achieve a visual dialog AI-dream to make a computer as efficient as a human in understanding, analyzing, and answering questions about a visual scene [1]. Providing an explanation for answering selection is essential in VQA [2]. Although significant progress has been made in QA models, VQA models still suffer from poor performance [3], [4]. The main reasons for this poor performance are as follows:

- The method that humans follow to solve problems differs from that used in VQA models. For example, while humans can easily recognize an older human in images, this could be harder for the model.
- Existing VQA models lack the ability to engage in higher-level reasoning [4]. For example, a question on tumor types according to specific properties, such as size, shape, and texture, requires more than object detection. When asking about a tumor larger than 5 mm, the model must detect all tumors in the image and all sizes of existing tumors, make a comparison, and answer the question.
- Many studies do not focus on the deep relationships between the expressed ideas in the text and image contents [5]. Many studies do not show whether the results are based on correct reasoning or coincidental answers [4].

One gap in this field is the requirement of a large, rich dataset with images and simple and complex question-answer pairs and their correlations, with no biased data for a specific subject. Therefore, much information was generated in the last three years [6]. Vu et al. [7] generated three medical datasets with complex questions.

Besides the complex question in the medical VQA gap, there are limitations related to this field. Although recent data has been made available by ImageCLEF-Med yearly and researchers enhanced existing data or generated new data, the information still has limitations and is insufficient in developing a robust and practical model used in the real world [8]. For example, a limitation appears in data size, which needs to be large to handle various questions and answers. Data with insufficient information about the images or patient history limit the real-world medical VQA agent system [8]. Although Kovaleva et al. [9] proposed patient history data, they extracted the history based on only one sentence. Furthermore, unbalanced or biased data are two other data shortcomings. The data is simple, with no complex questions, leading to a simple model that cannot answer complex problems. Besides, automatic data generation methods create a robust data problem [10]. All those data matters sufficiently affect the performance of VQA models. Researchers proposed different solutions and multi-models to exceed those borders and enhance the overall performance, which is still considered low.

Medical images, such as CT, MRI, Mammogram, and ultrasound, are affected during acquisition and transmission. These noisy images require a robust model that can pass this noisy limitation [11]. Although researchers, such as Nguyen et al. [12] and Zhan et al. [13], proposed models to exceed this problem and significantly enhance performance, the performance levels remain low. Besides those problems in medical VQA, pre-trained models, such as VGGNet and ResNet, which are regularly used, have fixed input image sizes that can affect the models' performance and make disease features invisible [14]. Finding an effective augmentation method can help in this case.

The medical VQA model can play a role in the medical VQA agent system to help patients understand their X-ray, CT, or MRI images. Medical VQA also helps students in the medical field. The accurate models help doctors and ray specialists acquire more information by asking questions about ambiguous objects in the image.

Medical VQA is a new and underexplored AI field. Researchers designed various multi-model VQAs to improve performance. These multi-models require further study to detect their pros and cons and overcome limitations. An in-depth analysis of the most recent models is needed to select a novel model structure that significantly enhances performance. Several surveys have focused on VQA in general field [1], [15], [16], [17], [18], [19], [20], [21], [22], [23] except Lin et al. [24], which is classified as the first survey in medical VQA. Table 1 contains the relative surveys.

Table 1 shows the published survey research in the VQA field from 2017 to 2023. Although many surveys were published, which denoted how active this field is, most of these publications were in the general field and did not have sections on medical VQA, their pros and cons, or the open challenges in the medical field except Lin et al. [24], Noor Mohammed and Srinivasan [26], and Lin et al. [27]. Lin et al. [24] and Mohamed and Srinivasan [26] propose a VQA survey study in the medical field. Those survey studies survey the methods designed for the 2018, 2019, and 2020 ImageCLEF challenges and two extra models on the VQA-RAD dataset. The Lin et al. [24] study also surveys the public medical VQA datasets except VQA-Med 2021, whereas Mohammed and Srinivasan [26] added VQA-Med 2021 and Diabetic Macular Edema (DME) dataset [28]. Additionally, they discussed the challenges and future studies in the field. Lin et al. [24] and Mohamed and Srinivasan [26] surveys, which are based on a comparison of 32 studies, require further work to cover and analyze more methods and datasets. The most recent survey, Lin et al. [27], is more comprehensive than the previous two surveys [24], [26] where it covers 44 studies with 47 models. The present survey study is an analytical review with 60 studies in medical VQA. It would be more comprehensive than the previous study. It proposes and compares more than 75 medical VQA models. Regarding a dataset, the existing reviews [24], [26], [27] surveyed eight, five, and eight datasets, respectively, whereas, in the proposed review, 16 datasets are analyzed. Moreover,

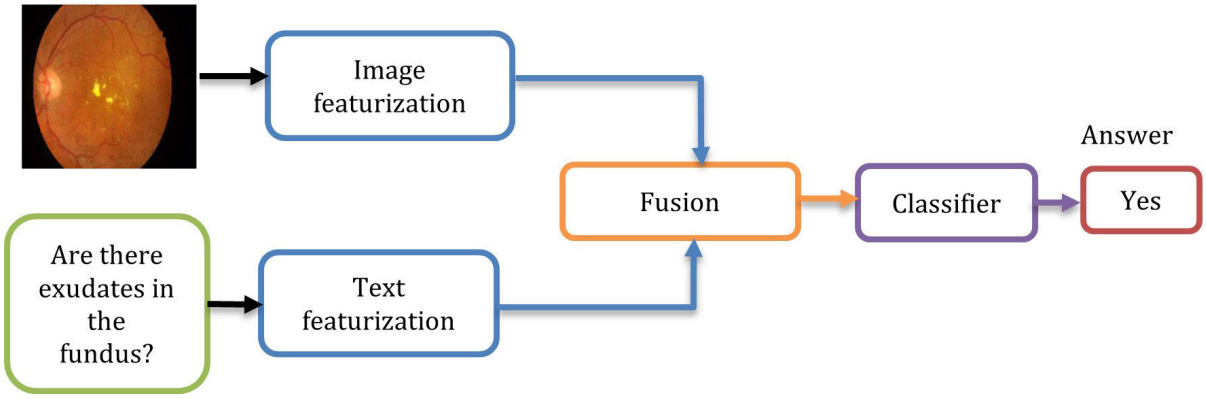


FIGURE 1. QA main architecture.

TABLE 1. A VQA state-of-the-arts researches.

Authors	Year	Paper
Wu et al. [15]	2017	Visual Question Answering: A Survey of Methods and Datasets.
Kafle and Kanan [16]	2017	Visual question answering: Datasets, algorithms, and future challenges.
Gupta [17]	2017	Survey of visual question answering: Datasets and techniques.
Zhang et al. [18]	2019	Information fusion in visual question answering: A Survey.
Katiyar and Wakode [25]	2020	A Survey On Visual Questioning Answering : Datasets, Approaches And Models
Manmadhan and Binsu [1]	2020	Visual question answering: a state-of-the-art review.
Patil and Patwardhan [19]	2020	Visual Question Generation: The State of the Art.
vrivastava et. al [20]	2020	Visual Question Answering Using Deep Learning: A Survey and Performance Analysis.
Zou and Xie [21]	2020	A Survey on VQA: Datasets and Approaches.
Sharma H. and Jalal A. [22]	2021	Asurvey of methods,datasets and evaluation metrics for visual question answering
Sahani ed al. [23]	2021	A Survey on Representation Learning in Visual Question Answerin.
Lin et al. [24]	2021	Medical Visual Question Answering: A Survey
Noor Mohamed and Srinivasan [26]	2023	A comprehensive interpretation for medical VQA: Datasets, techniques, and challenges
Lin et al [27]	2023	Medical visual question answering: A survey

this review study compares the techniques used in the medical field with those used in the general field, surveyed in the Sharma and Jalal study [22] in 2014-2020, which surveyed 80 models. In the present review, the statistical analysis for these studies in [22] is done to show which techniques are primarily used in the general field and check whether those influence the researchers in the medical field.

Furthermore, the Strengths, Weaknesses, Opportunities, and Threats (SWOT) analysis technique provides a clear view of a subject that helps the researchers understand what has already been done and their weaknesses. They also receive opportunities to consider new research about the threats they may face. This research has utilized SWOT to comprehensively analyze medical VQA datasets, techniques, attentions, and vision+language pre-trained models. The analytical study contributions are summarized as follows:

- Proposing a survey of existing medical datasets with their characteristics, generation, and statistical and SWOT analysis.
- Vision and text featurization techniques along with a fusion phase utilized in medical VQA undergo discussion in this survey. The statistical and SWOT analysis of these methods also take place.
- Statistical analysis of the text and vision featurization methods in general VQA, based on Sharma and Jalal [22], are compared with those in the medical field to check whether general VQA influences the researcher.
- We propose the challenges and give recommendations that may help the researcher start new research in the field.

The rest of this paper proposes the review assumptions and methodology in the second and third sections, followed by the VQA question types in the fourth section. Surveys of medical benchmark datasets have been proposed in the fifth section, followed by VQA evaluation metrics in the sixth section. Section seven proposes the medical VQA systems, and a discussion and statistical and SWOT analysis are discussed in the subsequent section. Besides, the open challenges and recommendations for researchers were proposed in the same section. Finally, we conclude the review in the last section.

II. THE REVIEW ASSUMPTION

Several assumptions can be taken into account while conducting a critical evaluation based on statistics and SWOT analysis for the dataset and methods in the field of med-VQA. These assumptions could consist of:

- **Representative Dataset:** The chosen datasets for the study are presumed to be typical of the broader field of medical VQA and include various medical diseases, image kinds, and question categories. Methodological Consistency: The presumption that methodology used in various research within the field follows consistent guidelines to enable meaningful comparisons and analysis.
- **Data quality:** The dataset's validity and integrity are guaranteed by the assumption that the data used in the reviewed studies is accurate, reliable, and correctly annotated.
- **Generalizability:** The presumption that the results and conclusions drawn from the dataset analysis and methodologies used in the evaluated research can be generalized to a broader context and applied to additional medical VQA scenarios.
- **SWOT framework applicability:** Assumption that the SWOT analysis framework is a suitable and valuable tool for assessing the advantages and disadvantages of the dataset and the methods used in the med-VQA area.
- **Validity of Statistical Analysis:** The validity of the presented results is predicated on the assumption that the statistical analyses carried out in the examined research were adequately planned, carried out, and interpreted.
- **Publication Bias:** The presumption that the studies that have been evaluated are a relatively complete and unbiased sample of the literature currently available in the field of medical VQA, without a significant bias towards publishing only significant or positive findings.

III. THE REVIEW METHODOLOGY

This critical review aims to thoroughly study and analyze existing benchmarks, techniques, and models in the medical Visual Question Answering (med-VQA) field. The methodology of the review can be summarized as follows:

- Based on the contribution of studying the authors' inspiration by the VQA in the general field based on Sharma and Jalal [22] that focuses on studies published between 2014 and 2021, and since the field of Visual Question Answering (VQA) in the medical domain emerged in 2018, the most relevant studies published between 2018 and 2021 were included in our survey. In order to stay up to date, we also considered some studies published in 2022 and 2023. Additionally, we included all studies from the imageCLEF challenges conducted between 2018 and 2021. However, It is important to note that this survey does not encompass any papers from the med-VQA challenge organized by imageCLEF in 2022, as no such event occurred during that year. Furthermore, since the imageCLEF 2023 conference was held in September 2023, any papers presented at that conference will not be included in this survey.
- The review covers and analyzes all benchmarks used in med-VQA since 2018, providing insights into their generation methods, sizes, validation procedures, question types, image types, and limitations.
- Since one criterion of the comparison between models is a performance metric, and it is one of the VQA gaps, those metrics are discussed.
- The review explores VQA components and techniques to provide researchers with a clear overview of VQA before delving into med-VQA models.
- The models are classified into sections and discussed based on their methods. Statistical and SWOT analyses are conducted. The statistical analysis focuses on the frequency of each method used in the literature along with its performance on different datasets. Moreover, the analysis is performed for each dataset, considering the frequency of its usage and the best accuracy achieved.Furthermore, the VQA in the general field started three years before it was in the medical field. Therefore, a statistical analysis is conducted to study whether the researchers influence the researchers in the general field.
- The SWOT analysis addresses several key questions: What significant research aspects exist in the field that can contribute to significant progress? What are the limitations of existing research? What opportunities do these limitations present for researchers? Lastly, what aspects do researchers need to be aware of?
- Finally, the review concludes with a discussion of challenges in the med-VQA field and provides recommendations to guide future research.

This methodology ensures a comprehensive and systematic analysis of med-VQA benchmarks, techniques, and models and provides valuable insights for researchers in the field.

IV. MEDICAL VQA DATASETS

Many VQA datasets have been made publicly available. These datasets can be classified based on the image type into four categories: clip-art, natural, synthetic, and hybrid. Figure 2 shows an example of the first three image types. To the best of our knowledge, there are five of nine publicly medical datasets.

A. VQA DATASET GENERATION

Generating questions and answer pairs in natural languages based on images is a new process known as visual question generation (VQG) [29]. The primary motivation behind this task is to provide a large-scale dataset to create practical VQA agents [30], [31]. There are three methods for visual question-answer pairs: manual, automatic (VQG), and semi-automatic. Figure 3 shows the dataset generation types. Manual VQA dataset generation is based on specialists creating the question and answer pairs, such as VQA-RAD [10] and SLAKE [32] datasets. One dilemma related to this method is that the size of these datasets is relatively

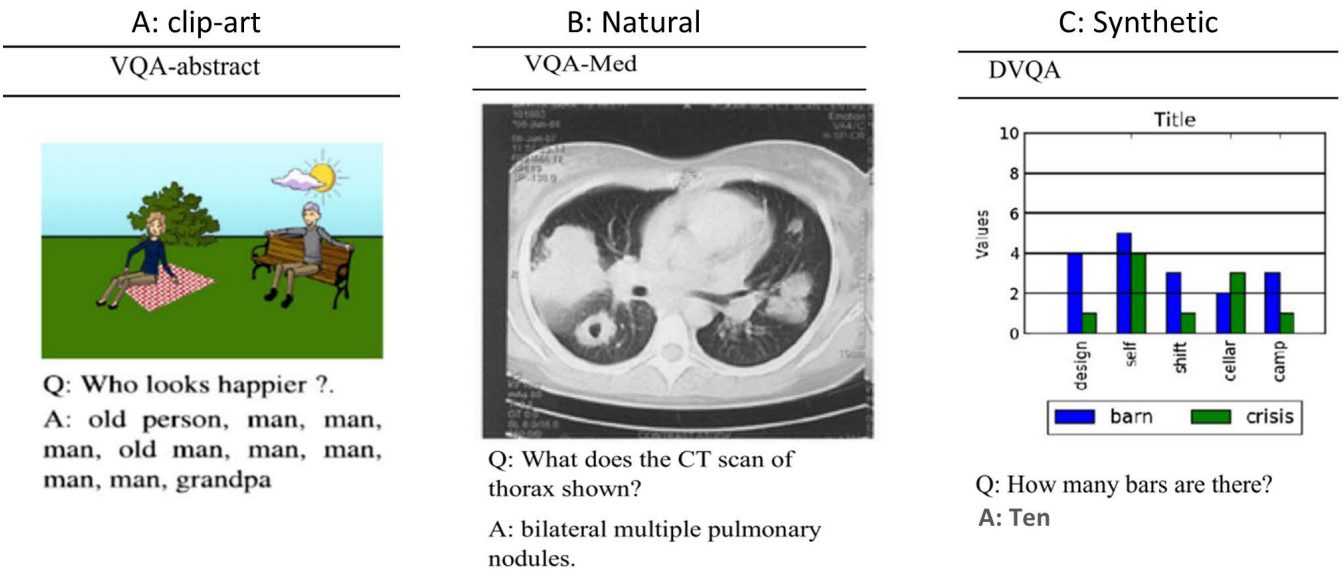


FIGURE 2. VQA question types.

small due to the lack of available specialists. Since VQA requires a large dataset, these datasets do not provide an efficient, practical VQA agent. Another type of VQA dataset generation is semi-automatic, based on automatic generation for question-answer pairs and authenticated by specialists, such as VQA-Med 2019. VQG refers to the automatic visual question-answer pair generation methods based on generating question-answer pairs with no human authentications. RedVisDial [9], Tools [7], BACH [7], and IDiRD [7] datasets are examples of VQG. The VQG datasets have two primary problems: noise and having no sense question-answer pairs [10], [33]. A generated VQG task is more intelligent and tricky than VQA because it requires deep background knowledge about the information related to the problem before designing the methods [30]. The VQG task has not undergone much exploration because of the effort required to explore it [30]. Even if it is possible to generate a dataset automatically from validated existing datasets, the medical dataset still suffers from size limitations. ImageCLEF included VQG tasks from challenges between 2019 and 2021. Sarrouti et al. proposed a comprehensive state-of-the-art survey for VQG [30]. Table 2 shows a comparison between the three dataset generation types in terms of size, authentication, errors, question sense, cost, and trust.

Eleven question types and four answer types (“yes” or “no”, numbers, categories, and locations) were used.

TABLE 2. Comparison between dataset generation types.

Criteria	Manually	Semi-automatic	Automatic
Size	small	small	large
Authenticated	yes	yes	no
Errors	low	medium	high
Question Sense	yes	yes	no
Cost (time/effort/money)	high	medium	low
Trusted	high	medium	low

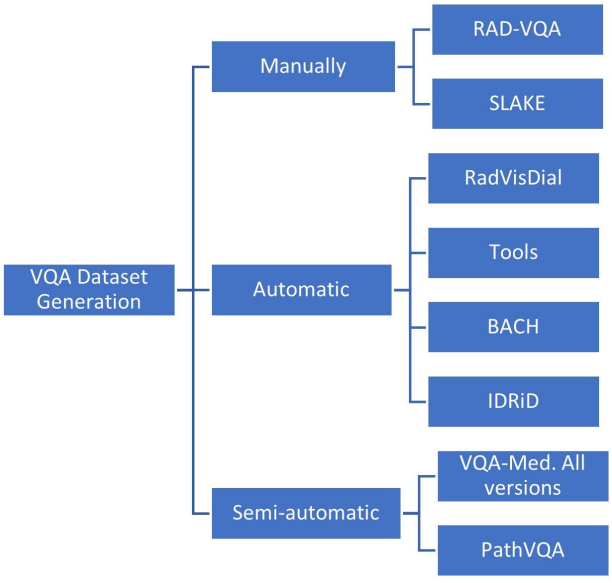
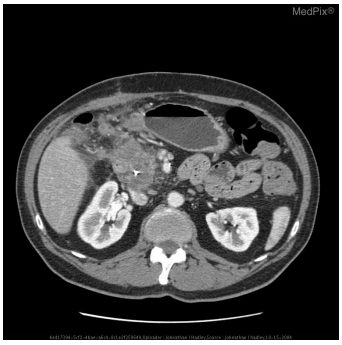


FIGURE 3. Dataset generation types.

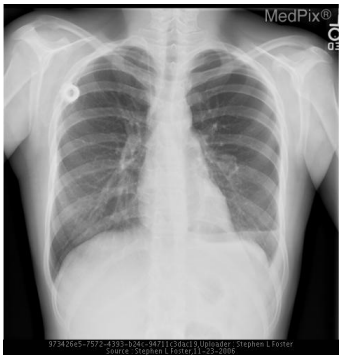
B. EXISTING VQA DATASETS

In 2018, the ImageCLEF-Med challenge [3] called on researchers for a medical VQA challenge. They provided the VQA-Med v1 dataset with 2,866 radiology images taken from PubMed Central articles and 6,413 question-answer pairs. These question-answers were generated automatically from corresponding image captions using the MS-COCO

(a) **Question:** Is this a KUB film?
Answer: MODALITY

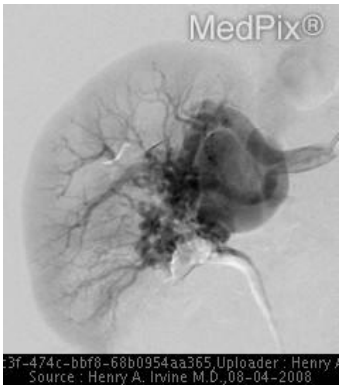


(b) **Question:** is the stomach filled?
Answer: yes

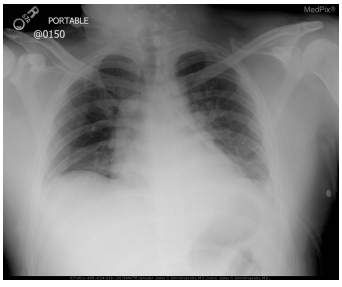


(c) **Question:** How was this film taken?
Answer: PA

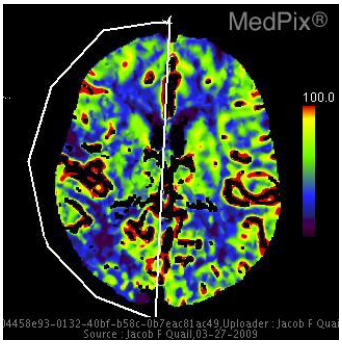
FIGURE 4. VQA-RAD dataset.



(a) **Question:** what kind of image is this?
Answer: an-angiogram



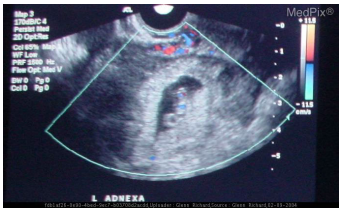
(b) **Question:** what is the plane of the x-ray?
Answer: frontal



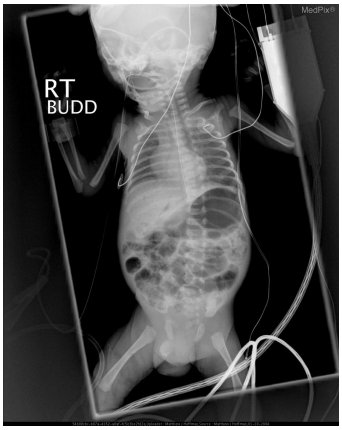
(c) **Question:** which organ system is shown in the ct scan?
Answer: skull and contents

FIGURE 5. VQA-Med 2019 dataset.

(a) **Question :** are there abnormalities in this mri?
Answer: no



(b) **question:** Question: what abnormality is seen in the image?
Answer: ectopic pregnancy

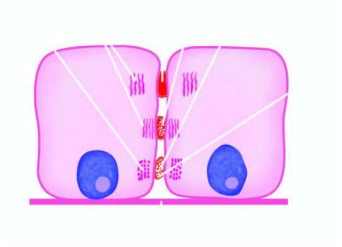


(c) **Question:** what is the primary abnormality in this image??
Answer: necrotizing enterocolitis

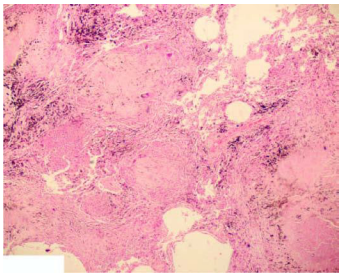
FIGURE 6. VQA-Med 2020 dataset.

dataset [34]. Therefore, some questions did not make sense. This dataset suffers from bias. In the same year, the VQA-RAD [10] dataset was made available publicly. The VQA-RAD was the first manual dataset included questions

(a) **Question:** How is a liver studded ?
Answers: with metastatic cancer

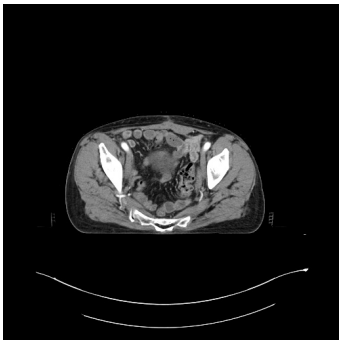


(b) **Question:** Are apoptotic cells in colonic epithelium shown ?
Answers: yesl

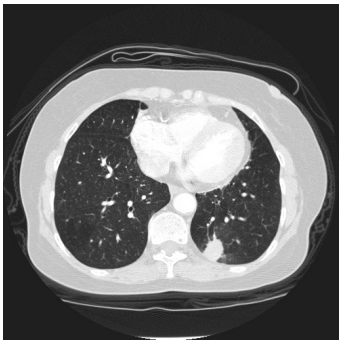


(c) **Qestion:** What are predominance of blastemal morphology and diffuse anaplasia associated with ?
Answers: specific molecular lesions

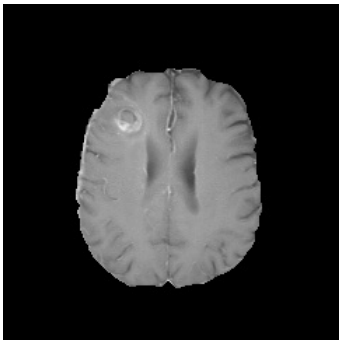
FIGURE 7. PathVQA dataset.



(a) **Question:** Which type of modality is shown about this image, MRI, CT or X-Ray?
Answer: CT

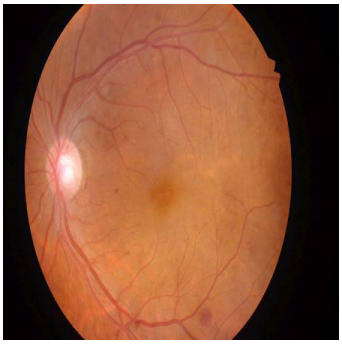


(b) **Question:** What type of medical image is this?
Answer: MRI



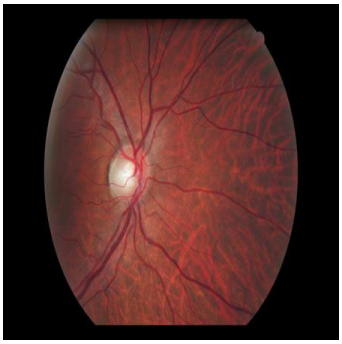
(c) **Question:** What modality is used to take this image?
Answer: CT

FIGURE 8. SLAKE dataset.



(a) **Question:** Are there hard exudate in the fovea?
Answer: yes

(b) **Question:** What what is the diabetic macular edema grade for this image?
Answer: 2



(c) **Question:** Are there hard exudate in this image?
Answer: no

FIGURE 9. DME dataset.

answered by clinicians. It has 315 radiology anatomical medical images; thus, one limitation of this dataset is its small size. In 2019 and 2020, ImageCLEF-Med provided the VQA-Med v2 [31] and VQA-Med v3 [35] datasets containing 4,200 radiology images with 14,292 question-answer pairs and 5,000 radiology images with 5,000 question-answer

pairs, respectively. All VQA-Med dataset versions relate to general medicine. The reset medical VQA datasets created in 2020 relate to specific specializations.

The PATHVQA [8], RadVisDial [9], BACH [7], Tools, and IDRiD [7] datasets deal with pathology, chest x-rays, breast cancer histology, surgical tools, and diabetic retinopathy specializations, respectively. The PATHVQA dataset has 4,998 pathology images with 32,799 open-ended questions. One problem of the PATHVQA dataset is that the lack of diversity and robustness in question-answer pairs created from captions using the linguistic rules method [8]. The RadVisDial [9] dataset contains 91,060 x-ray images with 455,300 question-answer pairs. These are split into 77,205, 7,340, and 6,515 images for training, validation, and tests. The RadVisDial [9] is the largest dataset available and is unique as it provides external information. The IDRiD dataset [36] contains 516 retina color fundus images with 220,000 question-answer pairs. The BACH dataset contains 420 microscopy images for breast cancer with 360 question-answer pairs. The Tools dataset includes data for seven surgical tools: grasper, hook, bipolar, scissors, irrigator, clip applier, and specimen bag. This dataset contains 2,523 images with one million question-answer pairs. The IDRiD, Tools, and BACH datasets contain complex questions. However, these three datasets use dataset annotation as the generation method for QA, which means the possibility of error if the original dataset annotation has an error.

Another manual medical VQA dataset is a semantically-labeled knowledge-enhanced (SLAKE) dataset, which is created in 2021 [32]. As VQA-RAD dataset, SLAKE is based on expertise humans for form the question answers pairs, but it is larger than VQA-RAD dataset and answer more composed and complex questions, including queries such as disease-causing, organ functionality, or disease treatment. SLAKE is a public medical bilingual dataset that has English and Chines question answers pairs. It covers human parts more than the previous existed datasets. It contains 642 radiology images with 14K question answers pairs. It covers 12 diseases on 39 human parts.

The medical datasets in 2022 are OVQA dataset [37], Diabetic Macular Edema (DME) dataset [28], EndoVis-18-VQA [38], and Cholec80-VQA [38].

OVQA dataset [37] has been created based on hospital FQAs. Physicians verified the template of the questions and answers. OVQA has 19,020 question-answer pairs about abnormality, modality, organ, plane, condition presence, and attribute others. The dataset is split into training, validation, and testing datasets with 2,000, 1,235, and 1,234 images related to 15,216, 1,902, and 1,902 question-answer, respectively. All the questions are about two modalities: x-ray and CT, which cover six body parts: hand, leg, head, and chest. The dataset includes 2001 images distributed in 70% are CT, and the remains are X-ray images.

Diabetic Macular Edema (DME) dataset [28] that is generated automatically from the Indian Diabetic Retinopathy Image Dataset (IDRiD) [36] and the e-Ophta dataset [39] is used. The dataset has 679 images with 13470 question-answer pairs distributed into 433,112, 134 images with 9779,2380, and 1311 for the train, validation, and testing dataset. The dataset has questions about exudates' grades. The dataset has specific questions with five answers. The questions have been assigned to a region or a whole image. It is classified as a manually generated dataset.

EndoVis-18-VQA dataset was generated by extracting images from the MICCAI Endoscopic Vision Challenge 2018 [40] dataset. Each image has two question types: one with a single-word answer (EndoVis-18-VQA (C)) and another with a sentence answer (EndoVis-18-VQA (S)). The question answers were generated based on the tissue, tool, interaction annotations, and bounding box used for tool-tissue interaction detection tasks [41]. Both versions have 1,560 images with 9,014 question-answer pairs and 447 images with 2,769 question-answer pairs for training and testing datasets, respectively.

Cholec80-VQA dataset has 21591 images generated from sampling 40 video sequences of the Cholec80 dataset [42] at 0.25 fps. Each image related to 2 questions using the phase annotations and tool-operation provided in the original dataset [42]. The dataset has two parts; the first is for classifying 14 single words (Cholec80-VQA (C)), and the second one is for sentence answers (Cholec80-VQA (S)). Each part of the dataset has 17,000 images with 34,000 question-answer pairs and 4,500 images with 17,000 question-answer pairs for training and testing datasets.

The newest medical dataset is a Patient-oriented Visual Question Answering (P-VQA) [43], which was published in 2023. The dataset contains 2,169 X-ray, CT, MRI, and Ultrasound images collected from hospitals with 24,800 question-answer pairs for 20 diseases of 12 body parts. The p-VQA dataset is provided with a knowledge graph showing the 13 relationship types between attributes and diseases. Those relations are built based on patient questions. The question-answer pairs are created based on the knowledge graph and templates, which are written manually. The dataset has 12 question types: symptoms, organs, diseases, therapy advice, medicine, treatment, examination items, prognosis, department, examination advice, prevention, pathogenesis, and review time. The dataset splitting is 1,526 images with 17,336 question-answer pairs, 218 images with 2,575 question-answer pairs, and 425 images with 4,889 question-answer pairs for taking, validation, and testing datasets, respectively. Figures 4-9 show examples from above datasets. Table 3 shows a summary of the described datasets.

V. EVALUATION METRICS

There are two types of questions: multiple-choice and open-ended questions. In multiple-choice, there is only one correct answer. Therefore, metrics like accuracy, recall, and precision can give a correct evaluation, but these performance metrics are not precision metrics in the open-ended question due to paraphrasing and synonyms. Therefore, other performance metrics are used in open-ended questions in VQA: Wu-

TABLE 3. Medical VQA benchmarks.

Beginning of Table					
Dataset/ Year	Size	Generation	Limitation	Ques Type	Public
VQA-RAD [10] 2018	315 images with 3,515 question-answer pairs	Manually	-small size -does not have composed and complex questions	-Modality -Plane -Color -Size -Attribute Other -Counting -Organ system -Abnormality -Object/ Condition presence -Other	✓
VQA-Med [3] 2018	2,866 images, 6,413 QA Training: 2,278 images 5,413 QA Validation: 324 images 500 QA.Testing: 264 images, 500 QA	Semi-automatic	-biased -does not have complex questions -unbalanced -the questions do not always make sense	-Location -Finding -Yes/No -Other	
VQA-Med [31] 2019	Training: 3,200 images, 12,792 QA. Validation: 500 images, 2,000 QA. Testing: 500 images, 500 QA.	Semi-automatic	-unbalanced classes -general and not suitable for specializations -too small to use in a real-world system -does not have complex questions	Modality -Organ system -Plan -Abnormality	✓
VQA-MED [35] 2020	Training: 4,000 images, 4,000 QA. Validation (500 images, 500 QA. Testing: 500 images, 500 QA.	Semi-automatic	-unbalanced classes -general and not suitable for specializations -too small to use in a real-world system	-Abnormality	✓
PathVQA [8] 2020	4,998 with 32,799 open-ended QA pairs	Semi-automatic	-need to add patient medical history and demographics with images to the dataset match exact ABP tests -insufficient diversity and robustness of question/ answer pairs created from captions by using the linguistic rules method -for the Path-VQA performance gap, although Convolutional Neural Network (CNN) and transfer learning methods are used, performance is still low	-Shape -Color -Location -Appearance -etc	✓
RadVisDial [9] 2020	91,060 images with 5 random questions for each image. Training: 77,205 images Validation: 7,340 images Testing: 6,515 images	Automatic	-does not support complex questions. -limited to abnormality questions with four answer choices -patient history is created from only one sentence -unbalanced data	-Abnormality	
BACH [7] 2020	516 images with 360,000 question-answer pairs	Automatic	-the generation method used for QA is based on dataset annotation, which raises the possibility of error if the original dataset annotation has an error -unbalanced data	-yes/no -numbers -categories	
IDRiD [7] 2020	420 images with 220,000 question-answer pairs	Automatic	-generation method used for QA is based on dataset annotation, which means the possibility of error if the original dataset annotation has an error	-yes/no -yes/no/Na	

TABLE 3. *(Continued.)* **Medical VQA benchmarks.**

Dataset/ Year	Size	Generation	Limitation	Ques.Type	Public
Tools [7] 2020	2,532 video frames with one million question-answer pairs.	Automatic	-generation method used for QA is based on dataset annotation, which means the possibility of error if the original dataset annotation has an error -unbalanced data	-yes/no -numbers -locations	
SLAKE [32] 2021	642 images with 14K question-answer pairs	Manually	-relatively small	-Shape -Size -Plane -Color -Quality -Abnormality -Modality	✓
OVQA [37] 2022	2,001 with 19,020 question-answer pairs. Training: 2000 images, 15,216 QA. Validation: 1,235 images, 1,902 QA. Testing: 1,234 images, 1,902 QA	semi-automatic	-unbalances classes -most image about hand -Abnormality -Modality -Organ -Plane -Condition presence -Attribute others		
DME [28] 2022	679 images, 13470 question-answer pairs. Train: 433 images 9,779 QA. Validation: 112 images, 2,380 QA. Testing: 134 images 1,311 QA.	Manually	-low diversity -unbalanced class -few questions	-exudates' grades	✓
EndoVis-18-VQA [38] 2022	Train: 1560 images 9,014: QA. Test: 447 images 2,769 QA	Automatic	-less annotated data -no information about question types or answer types which could be used to reduce the question complexity	-	✓
Cholec80-VQA [38] 2022	Train: 17K images 34K QA. Test: 4.5K ,9K QA.	Automatic	-less annotated data -no information about question types or answer types which could be used to reduce the question complexity	-	✓
P-VQA [43] 2023	2,169 images. Train: 1,526 images, 17,336 QA. Validation: 218 images, 2,575 QA Test: 425 images, 4,889 QA	Semi-automatic	-unbalanced	-symptoms -organs -diseases -therapy advice -medicine -treatment -examination items -prognosis -department -examination advice -prevention -pathogenesis -review time	✓

Palmer Similarity (WUPS) [44], Word-based Semantic Similarity (WBSS) [45], [46], BiLingual Evaluation Understudy (BLEU) [47], Concept-based Semantic Similarity (CBSS) [46], mean-per-type (MPT), and Metric for Evaluation of Translation with Explicit ORdering (METEOR) [47].details about each metric are discussed below.

- **Accuracy:** Accuracy is the ratio of the correct predicted answers to the number of all samples.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \quad (1)$$

- **Recall/ Sensitivity:** Recall or Sensitivity denotes to the ratio of the number of the correct predicted positive answers to all actual positive answers.

$$Recall/Sensitivity = \frac{TP}{TP + FN}. \quad (2)$$

- **Specificity:** Specificity is the ratio of the correct predicted negative answers to the number of all actual negative answers.

$$Specificity = \frac{TN}{TN + FP}. \quad (3)$$

- **Precision:** Precision is the ratio of the correct predicted positive answers to the number of all predicted positive answers.

$$Precision = \frac{TP}{TP + FP}. \quad (4)$$

- **F1-Score:**

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (5)$$

- **WUPS:** This metric is based on the semantic meaning and how much the actual answer differs from the predicted answer. The decision of which the predicted answer is true or false is controlled using a threshold. WUPS is calculated based on the following equation:

$$WUPS(a, t) = \frac{1}{N}\sum_{i=1}^{N} min\left\{ \prod_{a \in N} \max_{t \in N}(a, t), \prod_{t \in T^i} \max_{a \in T^i} \right\}.100 \quad (6)$$

where N, A, T denoted to the total number of questions, predicted answers, and the actual answers, respectively. $WUP(a, t)$ returns the position of words a and t in the taxonomy relative to the position of the Least Common Subsumer (a, t)

- **WBSS:** WBSS metric is based on finding the score of word similarity between the actual answer and the predicted answer. WBSS is calculated based on the following equation:

$$\begin{aligned} S(q, c) &= \text{softmax}(W_2 \max(\text{ReLU}(W_1 U), 0)) \\ U &= [\overrightarrow{U}; \overleftarrow{U}; \overrightarrow{U} \odot \overleftarrow{U}; \overrightarrow{U} - \overleftarrow{U}] \\ \overrightarrow{U}_i &= \text{GRU}(\overrightarrow{U}_{i-1}, x_i) \\ \overleftarrow{U}_i &= \text{GRU}(\overleftarrow{U}_{i+1}, x_i) \\ x_i &= [\text{BERT}(q); \text{BERT}(c_i)] \end{aligned} \quad (7)$$

where q and c are the question and context. $S(q, c)$ represent the similarity score between q and c. U, $\overrightarrow{U}$, $\overleftarrow{U}$ are the input embedded, forward hidden states, and backward hidden states, respectively. x_i is the input embeddings created by concatenating the BERT embeddings of the question q and the i-th context token c_i.

- **BLEU:** BLEU metric depends on analyzing the n-grams co-occurrences between the actual answer and the predicted answer. BLEU is calculated based on the following equation:

$$BLEU = BP.\exp\left(\sum_{n=1}^{N} W_n \log P_n\right) \quad (8)$$

where BP, W, and P denoted to Brevity Penalty, Positive weights summing to one, and entire corpus Precision score, respectively.

- **CBSS:** It is almost as same as WBSS metric, excep using MetaMap13 for biomedical concept extraction instead of tokenizing the ground truth and system-generated answers into words.
- **MPT:** This performance metrics was proposed solve the problem of the dataset with unbalanced distribution of question-type or bias answer distribution for each type of question. It is based on calculating the harmonic or arithmetic mean accuracy for each question type. It is calculated suing the following equation.

$$MPT = \sum_{t=1}^{T} A_t \Big/ T \quad or \quad MPT = T \Big/ \sum_{t=1}^{T} A_t^{-1} \quad (9)$$

where T and A denoted to the number of question types and Accuracy over question type t, respectively.

- **METEOR:** This metric aims to find the similarity by when align the words in the ground truth with predicted answers one to one. Not always such alignemt is found. the following equation is used for calculate METEOR.

$$METEOR = (1 - Pen) * F_mean \quad (10)$$

VI. VQA SYSTEM COMPONENTS

As noted, VQA consists of four components: image featurization, text featurization, fusion (joint) that combines image featurization and text featurization, classifier, and V+L pre-trained model. For statistical analysis in this section, we compared methods in the general field, and those in the medical field based on our state-of-the-art and Sharma and Jalal [22] VQA survey in the general field.

A. IMAGE FEATURIZATION

To easily apply mathematical operations to an image, the image is represented as a numerical vector called image featurization. There are various methods to calculate image featurization, such as scale-invariant feature transform (SIFT) [48], simple RGB vector, a histogram of oriented gradients (HOG) [49], and Haar transform [50]. In deep learning systems, such as CNNs, image featurization vectors are learned by the neural network. There are two choices for using deep learning: training a model from scratch or using a pre-trained model. The first method requires highly specific

computational sources and a large amount of data. As such, pre-trained models such as AlexNet [51], VGGNet [52], GoogLeNet [53], and ResNet [54][14] have been widely used in VQA. The most frequently used pre-trained model in VQA is ResNet because of the reasonable cost of its computational resources [1]. Ensemble of some deep learning method may be used too [55], [56], [57].

The visual features are extracted in this phase, and the most recent VQA image models are based on pre-trained CNNs, such as ResNet [58], [59], [60], DenseNet-121 [9], and VGGNet [18], [46], [61], [62]. Other methods used for image feature extraction that pay attention to questions in VQA include the Multi-modal Low-rank Bilinear (MLB) [59], Multimodal Compact Bilinear Pooling (MCB) [58], Global Multimodal Low-rank Bilinear (G-MLB) [7], and Multimodal Tucker Fusion for Visual Question Answering (MUTAN) methods [60].

B. TEXT FEATURIZATION

The text featurization phase is responsible for selecting and extracting features from the question and representing them as numerical vectors using word-embedding methods to apply mathematical processing. There are three word-embedding categories, namely count-based methods, prediction-based methods, and hybrid methods. One problem in word-embedding is choosing a suitable method for a given problem, which depends on a trial-and-error approach [1]. Count-based methods count the occurrence of words in the text using one-hot encoding, a co-occurrence matrix [63], and singular value decomposition (SVD) [64]. In prediction-based methods, word representation is learned based on a model. Neural network models [65], continuous bag-of-words (CBOW) [66], skip-gram [67] (which was used by Google as open-source and called word2vec) [68], long-sort-term memory (LSTM) [69], gated recurrent units (GRU) [70], skeleton-based [55] are examples of prediction-based word embedding. Hybrid methods are created from count-based and prediction-based methods. The global vectors (Glove) method proposed by Pennington et al. [71] is an example of a hybrid method. The most common methods used in question models are LSTM [8], [9], [72], GRU [8], [72], RNNs [46], [73], [74], [75], Faster-RNN [8], [72], and the encoder-decoder method [7], [58], [59], [60], [61], [76]. In addition to the previous methods, pre-trained models have been used, such as Generalized Autoregressive Pretraining for Language Understanding (XLNet) [77] and the BERT model [61], [78]. Some models have ignored text featurization and convert the problem into image classification problem [57], [79], [80]

C. FUSION METHODS

Since both text and image featurization are independently processed, a fusion of the two is required for generation answers. There are three fusion types: baseline fusion models, end-to-end neural network models, and joint attention models [1]. In baseline fusions, various methods are used, such as concatenation [81], element-wise multiplication, element-wise addition [82], all of them [83], or a hybrid of these methods with a polynomial function [84]. End-to-end neural network models can be used to fuse image and text featurization. Various methods are currently used, including neural module networks (NMNs) [85], multimodal, MCB [58], dynamic parameter prediction networks (DPPNs) [86], multimodal residual network (MRNs) [87], cross-modal multistep fusion (CMF) networks [88], basic MCB model with a deep attention neural tensor network (DA-NTN) module [89], multi-layer perceptron (MLP) [90], and encoder-decoder method [91], [92]. The main reason for using the joint attention model is to address the semantic relationship between text attention and question attention [1]. There are various joint attention models, such as the word-to-region attention network (WRAN) [93], co-attention [94], the question-guided attention map (QAM) [95], and question type-guided attention (QTA) [96].

Neural network methods such as LSTM and encoder-decoder are also used in the fusion phase. Verma and Ramachandran [61] designed a multi-model that used encoder-decoder, LSTM, and GloVe.

VQA attention scheme: Attention is utilized to identify semantic features in questions, images, or both. It improves the interaction between questions and visual features by focusing on specific words in the question and connecting them with specific regions or objects in the image. Attention mechanisms can be classified into single-hop and multi-hope attentions, based on the attention layers number [22]. Shih et al. [97] developed an attention-based approach for VQA that has recently emerged as a key element in almost all architectures. Current strands of research cover co-attention architectures for the generation of simultaneous attention in both textual and visual modalities, which heightens the accuracy of predictions [81], [94]. Nevertheless, an important challenge with global co-attention mechanisms relates to their limited ability to model interactions and attention among individual image regions and text segments, such as at the level of the word token. Dense co-attention networks, including BAN [98] and DCN [99], have been developed to solve this challenge, wherein every image region can interact with any - and every - word in the question. Resultantly, models of this kind can produce a more refined understanding and reason regarding the image-question relationships; as such, VQA performance increases. Despite this, the absence of self-attention within every modality in dense co-attention networks such as BAN and DCN is a bottleneck, such as word-to-word relationships in the question and region-to-region relationships in the image [100]. Yu et al. [100] developed a deep Modular Co-Attention Network (MCAN) to solve this bottleneck, which is composed of various Modular Co-Attention (MCA) layers. An MCA layer, in turn, contains two general attention units: guided attention (GA) and self-attention (SA). In the case of the latter, SA, it captures intra-model interactions such

as word-to-word and region-to-region interactions, whereas GA captures cross-model interactions such as word-to-region and region-to-word, which is achieved using a multi-head attention architecture. Although it is flexible and expressive, there are problematic aspects to this type of attention; in particular, the result is always a weighted combination of value pairs among which the model is attending. This can create challenges when closely related context is unavailable for the model to attend over, such as when there is a word with no corresponding image region or context word. In cases such as this, attention would lead to excessive noise or, more severely, the distraction of the output vector, which can undermine performance. Building on Huang et al. [101], Rahman et al. used the Attention on Attention (AoA) module to resolve this limitation. Cascading of the AoA module occurs multiple times to produce a new Modular Co-Attention on Attention Network (MCAoAN), which is an enhanced version of MCAN [100]. Through two independent linear transformations [101], similar to GLU [102], the AoA module produces both an information vector and an attention gate. To generate an information vector, the query context is concatenated with the attention results and a linear transformation is applied. Ben-Younes et al. designed MUTAN [60], which uses multi-modal tensor-based Tucker decomposition to parametrize the interactions of bilinear between question and image features. Minh et al. utilized the operation of an inner product instead of applying low-rank bilinear pooling, designing a full-rank bilinear transformation G-MLB [103] to obtain significant answer ranges. All previous approaches gave equal importance to question and image features. Vu et al. claimed that paying more attention to the question would enhance the results. Therefore, they designed a Question-Centric Multi-modal Low-rank Bilinear (QC-MLB) approach [7].

Various attention schemes are used in VQA models. Many models have used SAN [9], [10], [12], [46], Bilinear Attention Network (BAN) [12], MCB [10], [46], and MFB [61] attentions. In 2020, Kovaleva et al. [9] applied two attention architectures in the models: Late Fusion Network (LF) and Recursive Visual Attention Network (RVA).

D. ANSWER CLASSIFICATION AND GENERATION

This phase is responsible for producing the answer. Most researchers designed classification VQA models due to easiness, whereas others designed answer generation models. Several methods were used, such as a Softmax layer for classification and LSTM or CNN models for a generation.

E. VISION-AND-LANGUAGE PRE-TRAINED MODEL

ResNet [54],GoogLeNet [53], and VGG [52], among other models pre-trained on ImageNet [104], have contributed to significant early advancements in enabling diverse downstream CV tasks. For several NLP applications, there are some pre-trained models, such as XLNet [77], BERT [78], and RoBERTa [105], attained high accuracy results compared to the other transformer-based models in the state-of-the-arts. Multiple researchers have used transformer learning for the vision and language areas separately to resolve the challenge, seeking to use external data to pre-train image features before the training of feature fusion and generating a prediction [33]. Nevertheless, these studies have overlooked the degree to which the pre-trained features are applicable and compatible for cross-model fusion [33].

Driven by the usefulness and value of XLNET [77], BERT [78], and other large-scale pre-trained language models, recent researchers have sought to generate image-text joint embedding from pre-training transformer-based models on V+L datasets [92]. In turn, the joint embedding is fine-tuned with a set of V+L tasks, which has been shown to produce remarkable results. What distinguishes the models is their pre-training strategies and cross-modality architecture. Specifically, UNITER [106] and VisualBert [92] used a single stream of transformers to learn image-text embedding jointly. By contrast, LXMERT [107] and ViLBERT [108] incorporated a pair of separated transformer blocks on image and text input, along with a third fusion transformer block for cross-modality [92].

Active research in this area has proposed pre-trained V+L models [92], [106], [107], [108], [109], [110], [111], [112], [113], [114], [115], [116], [117], [118], [119], [120], [121], [122], [123], [124], [125], [126], [127], [128], [129], [130], [131], [132], [133], [134], [135], [136], [137], [138], [139], [140], [141], [142], [143], [144], [145] to learn V+L representations for specific V+L tasks. Nevertheless, almost all prior studies have not attempted to solve the problem of learning these representations through the explicit detangling of multi-modalities and the incorporation of visual concepts, and they are not capable of directly performing downstream generation tasks [146].

Pre-trained models like VGGNet or Bidirectional Encoder Representations from Transformers (BERT) that are trained for vision or NLP are not efficient for VQA models [147]. Other pre-trained models trained for vision and text datasets are more effective for VQA use. Examples of these pre-trained model are UNITER [106], LXMERT [107], VisualBERT [92], PixelBERT [114], and ClinicalBERT [148]. Li et al. [92] compared four V+L pre-trained models, namely VisualBERT [92], LXMERT [107], PixelBERT [114], UNITER [106], CTL [144], VLMixter [145], BLIP [149], OFA [150], CoCa [151], BEIT-3 [152], PaLI [153], and BLIP-2 [154]. They determined that the pre-trained model using VisualBERT achieved the highest AUC performance at 0.987, whereas the model using PixelBERT earned the lowest score. Table 4 shows the performance of the existing pre-trained models that are fine-tuned on the VQA-Med 2019 dataset and which model was utilized in the medical field.

VII. MEDICAL VQA MODELS

Medical VQA is an active topic where much research has occurred. In this section, we propose the VQA models

TABLE 4. Vision and language pre-trained.

Authors	Model	Year	text-dev	Test-std	Med VQA
Lu et al. [108]	ViLBERT	2019	70.55	70.92	
Li et al. [92]	VisualBERT	2019	75	75.23	✓
Tan et al. [107]	LXMERT	2019	72.42	72.52	✓
Zhou et al. [109]	Unified VLP	2019	70.5	70.7	
Su et al. [110]	VL-BERT	2020	71.79	72.22	
Guo et al. [111]	LAMP	2020	71.93	71.9	
Cho et al. [112]	X-LXMERT	2020	71.1	71.2	
Li et al. [113]	TDEN	2021	72.5	72.8	
Chen et al. [106]	UNITER	2020	73.82	74.02	✓
Huang et al. [114]	Pixel-BERT	2020	74.45	74.55	✓
Li et al. [115]	ERNIE-ViL	2020	74.95	75.1	
Zhang et al. [116]	DeVLBert	2020	71.1	71.5	
Guo et al. [111]	LAMP	2020	72.48	72.62	
Luo et al. [117]	CAPT	2020	72.78	73.03	
Li et al. [118]	UNIMO	2021	73.79		
Zhang et al. [119]	VinVL	2021	76.52	76.60	
Li et al. [120]	SemVLP	2021	74.52	74.68	
Kim et al. [125]	ViLT	2021	71.26		
Yang et al. [144]	TCL	2022	74.90	74.92	
Wang et al. [145]	VLMixer	2022	76.61	72.89	
Li et al. [149]	BLIP	2022	78.25	78.32	
Wang et al. [150]	OFA	2022	82.0	82.0	
Yu and Wang [151]	CoCa	2022	82.3	82.3	
Wang et al. [152]	BEIT-3	2023	84.19	84.03	
Chen et al. [153]	PaLI-17B	2022	84.3	84.3	
Li et al. [154]	BLIP-2	2033	82.19	82.30	

in the medical field based on those in ImageCLEF VQA challenges, CNN-LSTM-based models, image classification-based models, ensemble-based models, vision-and-language (V+L) transfer learning models, and models based on external knowledge.

A. IMAGECLEF VQA CHALLENGES

ImageCLEF calls for challenge yearly, where it started the first call in medical challenge in 2018 [3]. Although 28 groups were registered for this challenge, only five groups sent results in 17 runs [46], [73], [74], [75], [93].

Most groups built their models based on deep learning. RNN, such as BiLSTM and LSTM, were used for text featurization, whereas encoder-decoder-based frameworks, VGG, ResNet, and Inception-ResNet-v2, were used for vision featurization. SAN and MCB attentions were used by the NLM participant [46], whereas MFB and ETM were used by the UMMS participant [93]. Their models achieved the highest score of 0.162 and 0.186 for BLEU and WBSS, respectively. The NLM model achieved the highest WBSS performance with a score of 0.338.

Peng et al. [93] provided the winning model in the ImageCLEF 2018 challenge. The model was based on ResNet152 fine-tuned with ETM for vision, LSTM for text, and co-attention with MFH, followed by a convolution layer and ReLu layer for fusion. The models' performances in the first challenge were considered poor, unlike in the second challenge, where the performance levels improved [31].The best BLEU score was 0.644, whereas the best BLEU performance was 0.162 in 2018. This progress shows how this field encourages researchers to develop more robust medical VQA models. In this challenge, 17 teams [62], [103], [155], [156], [157], [158], [159], [160], [161], [162], [163], [164] submitted 90 runs [31]. Like the first challenge, the models based on deep learning used RNN and CNN with or without pre-trained models and focused on aligning the text with images. The winning Hanlin model was based on VGG-16 and global average boolling, BERT, and co-attention for visionm text and fusion phases, respectively [156].

Like the previous two challenges, ImageCLEF 2020 and ImageCLEF 2021 were based on deep learning [35], [165]. In the ImageCLEF 2020 challenge, 30, 11 [35], [61], [166], [167], [168], [169], [170], [171], and 62 teams for team registration, teams submitted, and runs, respectively. This challenge featured CNN, such as VGGNet and ResNet, transformers, such as RNN and BERT, multi-modal factorized bilinear (MFB) pooling, and multi-modal factorized high-order pooling (MFH) for vision, text, and fusion, respectively. Liao et al. [55] were the 2020 winners. They designed a method based on Skeleton-based Sentence Mapping (SSM). For the visual aspect, they made several models based on VGGNet, DenseNet, ResNet, NextNet, and mobileNet, and several ensemble models from all those visual parts. The winning model had an ensemble of all these models with different versions. For fusion, class-wise and task-wise were used. The best model with the ensemble of the last pre-trained vision models achieved 0.496 and 0.542 for accuracy and BLEU performances, respectively.

In ImageCLEF 2021, there were 48, 13 [57], [79], [80], [165], [172], [173], [174], [175], [176], and 68 registered teams, submitted teams, and runs, respectively, in the VQA task [165]. The vision part of most VQA multi-models is based on CNN models, such as ResNet, VGG, and

TABLE 5. ImageClef challenges information.

Year	2018	2019	2020	2021
Dataset	VQA v1	VQA v2	VQA v3	VQA v4
#Teams Registered	28	61	30	48
#Teams Submitted	5	17	11	13
# runs	17	90	62	68
Winner	Peng et al. [93]	Hanlin [156]	Liao et.al, (AIML) [55]	SYS-HCP [57]
BLEU	0.162	0.644	0.542	0.416
WBSS	0.0338			
Accuracy		0.624	0.496	0.382

TABLE 6. The models' comparison on the Med-VQA 2018.

Ref.	year	text	vision	fusion	Accuracy	BLEU	WBSS
Peng et. al [93]	2018	LSTM	ResNet-152	MFH		0.162	0.185
		ETM+LSTM	ResNet-152	MFH		0.158	0.186
		LSTM	ResNet-152	MFB		0.16	0.184
TU [74]	2028	BiLSTM	Inception ResNet v2	Att.		0.135	0.174
Abacha et al. [46]	2028	LSTM	VGG-16	SAN		0.121	0.174
Abacha et al. [46]	2028	LSTM	ResNet-152	MCB		0.085	0.144
JUST [75]	2028	LSTM	VGG-16	Concat		0.061	0.122
FSTT [73]	2028	BiLSTM	VGG-16	Concat		0.054	0.101
Gupta et al. [177]	2020	Glove+Bi-LSTM	Inception-Resnet-v2			0.132	0.162

TABLE 7. The models' comparison on the VQA-RAD 2018.

Ref.	year	text	vision	fusion	Accuracy	BLEU	WBSS
Lau et al. [10]	2018	LSTM	ResNet-152	MCB	0.43	0.3675	
Nguyen et al. [12]	2019	Glove	CNN ensemble	SAN	open-0.407 close 0.701		
				BAN	open-0.439 close 0.751		
Gupta et al. [177]	2020	Glove+Bi-LSTM	Inception-Resnet-v2			0.411	0.437
Zhan et al. [178]	2020	Glove+ GRU	GRU	BLOCK	open-0.6 close 0.79		
Gong et al. [57]	2021	LSTM	ResNet-31	CMSA	0.732		
Do et. al [56]	2021	LSTM	Ensemble	-	0.67		
Liu et al. [179]	2021	LSTM	ResNet	CNN	0.721		
Khare et al. [180]	2021	BERT	ResNet-152+GAP	multi-head att	0.72		
Dhanush et al. [181]	2021	BiLSTM	ResNet	CNN	open 0.678		
		BiLSTM	ResNet	CNN	close 0.721		
Cong et al. [182]	2022	BERT	ResNet-152	PCBI	0.758 $\pm$3.1		
Wang et.al. [183]	2022	BioWordVec and LSTM	Ensemble	attention	0.741		
Wang et.al. [184]	2022	LSTM	Ensemble	PCBI	0.688		

DenseNet, whereas text parts are based on LSTM and transformer-based, such as BioBERT and BERT. Fusion parts are based on pooling strategies, such as MFH and MFB. The winning SYSU-HCP model was based on ensemble learning with 0.416 and 0.382 for BLEU and accuracy, respectively [57], [165]. Their model is based on eight models: ResNeSt-50, ResNet-50, VGG-19, VGG-16, ResNeSt-50-HAGAP, ResNet-50-HAGAP, VGG-19-HAGAP, and VGG-16-HAGAP. They augmented the data using a mixup strategy during the training. Table 6 summarizes the ImageCLEF challenges from 2018 to 2021.

Al-Hadhrami et al.[2] is a fused of multiple of Al-Hadhrami et al.1 models based on the greedy soup technique.

B. CNN-LSTM BASED MODELS

Most models fall into the CNN-LSTM-based methods. These models aim to solve problems in the field, such as data limitation, required answer types, and text and vision reasoning. The answer to the medical question depends on the end-user [177]. When the end-user is a patient or student, the answer could be simple; for example, "yes" or "no" could be enough, but the answer should have more details when the end-user is a doctor or specialist. Therefore, Gupta et al. proposed a hierarchical multi-model that depends on segregating the question using the SVM traditional machine learning classification to adopt the answer to the end-user. Their model was based on using the Glove and

TABLE 8. Models' Com parison on the Med-VQA 2019.

Ref.	year	text	vision	fusion	Accuracy	BLEU	WBSS
Zhejiang [156]	2019	BERT	VGG-16	MFB with co-att	0.624	0.644	
minhvu [103]	2019	BERT	ResNet-152	MLB+ MUTAN	0.616	0.634	
TUAI [155]	2019	BERT	Inceprion-ResNet v2	concat	0.606	0.633	
UMMS [160]	2019	BiLSTM	ResNet-152	MFH with co-att	0.566	0.593	
IBM Research AI [185]	2019	LSTM	VGG-16	Att.	0.582	0.558	
LIST [163]	2019	LSTM	DenseNet-121	concat	0.556	0.583	
Turner.JCE [158]	2019	LSTM	VGG-19	concat	0.536	0.572	
JUST19 [164]	2019	non	VGG-16	non	0.534	0.591	
Team-PwC-Med [159]	2019	LSTM	ResNet-50	Att	0.488	0.534	
Techno [62]	2019	LSTM	VGG-16	concat	0.462	0.486	
Dear stranger [161]	2019	GRU	Xception	Att.	0.21	0.393	
abhishek-thanki [157]	2019	LSTM	VGG-19+ DenseNet	EWM[1]	0.16	0.461	
Bounaama et al. [62]	2019	LSTM	VGG-16	SAN+BAN	0.462	48.5	
A et al. [186]	2019	Non	CNN	Non	0.838		
		Non	VGG-16	Non-	0.837		
		Non	VGG-19	Non	0.8344		
			MobileNet	RNN	0.846		
Zhou et al. [109]	2020	BERT	VGG-16	non	0.938	0.912	
Ren and Zhou [14]	2020	non	ResNet-152		0.659	65.9	67.8
Vu et al. [7]	2020	Skip-though vector	ResNet-152	QC-MLP att.	0.6033		
Dhanush et al. [181]	2021	BiLSTM	MobileNet	SAN	0.808		
		BiLSTM	VGG-19	SAN	0.72		
		BiLSTM	VGG-16	SAN	0.72		
Sharma et al. [187]	2021	BERT	ResNet-152	MFB		0.605 $\pm$0.021	
Al-Sadi et al. [188]	2021	non	VGG		0.608	0.634	
Gasmi et al. [189]	2021	Bi-LSTM	effecientNet	attention	0.42		

TABLE 9. Models' comparison on the Med-VQA 2020.

Ref.	year	text	vision	fusion	Accuracy	BLEU
AIML [55]	2020	Skelton-based	ensemble CNN		0.496	0.542
TheInception-Team [171]	2020	Non	VGG-16	TCR+EWM[1]	0.48	0.511
bumjun-jung [169]	2020	BioBERT	VGG-16+GAP		0.466	0.502
HCP-MIC [170]	2020	BioBERT	BBN-ResNet-50	non	0.426	0.462
NLM [166]	2020	LSTM	ResNet-50	not mentioned	0.4	0.411
HARDENDRA-KV [61]	2020	BERT	VGG-16	non	0.378	0.439
Shengyan [168]	2020	Glove+ GRU	VGG-16	MFH co-att	0.376	0.412
kdevga [167]	2020	BERT	VGG-16	non	0.314	0.35
Noor Mohamed and Srinivasan[2] [26], [190]	2020	Glove+ LSTM	VGG-16	concate	0.282	0.33
Liao et. al [55]	2020	SSM	Ensemble	class-wize and task-wise	0.446	0.486
		SSM	Ensemble		0.494	0.539
		SSM	Ensemble		0.496	0.54
		SSM	Ensemble		0.496	0.542

TABLE 10. Models' comparison on the Med-VQA 2021.

Ref.	year	text	vision	fusion	Accuracy	BLEU
Gong et al. [57]	2021	non	Ensemble		0.382	0.416
Xiao et al. [172]	2021	Bio-BERT	VGG16+ GAP		0.362	0.402
Islami et al. [79]	2021	non	BBN+ ResNet		0.348	0.391
Li and Liu [173]	2021	Bio-BERT	VGG-8	non	0.316	0.352
Schilling et al. [80]	2021	Non	DenseNet-121	MFH	0.236	0.276
Li et al. [174]	2021	LSTM	ResNet-34	MFB + co-att	0.222	0.225
Noor M. and Srinivasan [175]	2021	LSTM	VGG-16	LSTM	0.196	0.221
Noor M. and Srinivasan [26]	2021	LSTM	VGG-16	LSTM	0.196	0.221

BiLSTM technique for text featurization and the Inception-Resnet-v2 pre-trained model for vision featurization. They concatenated those features using the concatenation layer followed by the batch normalization layer. They evaluated

TABLE 11. Models' comparison on PATHVQA dataset.

Ref.	year	text	vision	fusion	Accuracy	BLEU
He et al. [8]	2020	GRU	Faster-RCNN	MFB	(yes-no) 0.682	(open)0.324
		LSTM	ResNet-152	GLU	(yes-no) 0.576	(open) 0.133
		LSTM	VGG-16	BAN	(yes-no) 0.594	(open) 0.197
		LSTM	Fater-CNN	MCB	(yes-no) 0.62	(open) 0.212
		LSTM	ResNet-152	SAN	(yes-no) 0.601	(open) 0.198
Do et. al [56]	2021	LSTM	Ensemble		0.488	
Sharma et al. [187]	2021	BERT	ResNet-152	MFB	0.636±0.020	

TABLE 12. Models' comparison on the SLAKE dataset.

Ref.	year	text	vision	fusion	Accuracy
Liu et. al [32]	2021	LSTM	VGG-16		0.7273
		LSTM	FCN+ VGG-16	CMSA	0.7536
		LSTM	VGG-16	SAN	0.7427
		LSTM	VGG-16	BAN	0.382
Dhanush et al. [181]	2021	BiLSTM	ResNet	CNN	open 0.811 close 0.821
Cong et al. [182]	2022	BERT	ResNet-152	PCBI	0.825 ±0.
Haridas et. al. [191]	2022	BERT	ResNet-50 and DETR	VilBERT	0.824
Huang et. al. [43]	2023	BERT+ GRU	ResNet-50	Concat	0.806

TABLE 13. Models' comparison on the DME dataset.

Ref.	year	text	vision	fusion	Accuracy
Tascon et al. [28]	2022	LSTM	ResNet101	e multi-glimpse att	0.8345
Tascon et al. [28]	2023	LSTM	ResNet101	e multi-glimpse att	0.8459
Al-Hadhrami et al. [192]	2023	ELECTRA	SWIN	e concat.	0.85.74
Al-Hadhrami et al. [192]	2023	ELECTRA	SWIN	e concat.	0.87.41

TABLE 14. Models' comparison on several medical VQA datasets.

Ref.	year	text	vision	fusion	Dataset	Accuracy	BLEU	WBSS
Gupta et al. [177]	2020	Glove+Bi-LSTM	Inception-Resnet-v2		VQA-RAD 2018 + Med-VQA 2018		0.257	0.288
Kovaleva et al. [9]	2020	LSTM	DenseNet-121	SAN LF RVA	RadVisDail	used macro avg f1		
Vu et al. [7]	2020	Skip-though vector	ResNet-152	QC-MLP att.	BACH	0.9423±0.0027		
Vu et al. [7]	2020	Skip-though vector	ResNet-152	QC-MLP att.	Tools	0.9653±0.0004		
Vu et al. [7]	2020	Skip-though vector	ResNet-152	Multi-glimpse att.	IDRiD	0.9030 ±0.0034		
Huang et. al. [43]	2023	BERT+ GRU	ResNet-50	Concat	P-VQA	0.9545 ±0.13		

their model on two public datasets: RAD-VQA and VQA-Med v1. The model with segregated questions outperformed the model without segregation on both datasets and combined those two datasets. The model achieved 0.411, 0.132, 0.257 for BLEU on RAD, VQA-med, RAD-VQA+VQA-med datasets, respectively, and 0.437, 0.162, and 0.288 for WBSS on those three datasets, respectively. Their model did not outperform the Zhou et al. [74] model, which differed from their model thanks to the use of the fusion method. Zhou et al. [74] also used attention. Their model suffered from errors that they claimed could be returned to the auto-generation dataset or showed the unsuitability of the performance metric.

Using a pre-trained model before the fusion phase as a solution for limiting data problems enhances the accuracy but does not consider the alignment between text and image [33]. Therefore, Gong et al. proposed a cross-model self-attention (CMSA) to solve this problem and reformulated the pre-trained model using an external dataset to adopt it for a multi-task model of a multi-modality dataset. Their model

was based on ResNet-31 and a decoder with three MLP layers for the vision phase, LSTM for the text phase, and CMSA for the fusion. They used CMSA to focus on the image encoder for the representation learning instead of the fusion feature. The authors evaluated their proposed model on RAD-VQA and three external datasets: chest X-Ray 2, brain MRI [193] and abdominal CT 1. The model achieved 61.5%, 80.9%, and 76.3% accuracy for open-ended, closed-ended, and overall, respectively.

Most VQA learn single reasoning for fusion representation, which is more suitable for closed-end questions than open-ended questions. Zhan et al. built a multi-model that learns the different reasoning representations for each TCR and QCR question type [13]. Their model is a multi-level reasoning skill suitable for tasks of complex medical VQA. Their model is based on the Nguyen model [12] as a backbone. They evaluated their reasoning modules on the VQA-RAD dataset and achieved 60%, 79.3%, and 71.6% accuracy for open-ended, close-ended, and overall questions, respectively.

C. IMAGE-CLASSIFICATION-BASED MODELS

Although the visual questions are modality questions, some authors prefer to convert the problem to an image classification problem and use the answers as labels. This method faces a generalized problem since the question could change. Furthermore, the model cannot know whether a question is framed negatively. Although this is the case, three of seven groups in the ImageCLEF 2021 challenge used this method [57], [79], [80], and the owner model was based on this method [57]. Lubna et al. [186] proposed a VQA system based on modality questions produced by ImageCLEF 2019. They converted the problem into an image classification problem and applied four model structures, namely, VGG-16, VGG-19, MobileNet, and CNN, from scratch. Their models' accuracy achieved 0.838, 0.8344, 0.846, and 0.838 for VGG-16, VGG-19, MobileNet, and CNN, respectively.

D. ENSEMBLE-BASED MODELS

Nguyen et al. utilized MAML to solve the data limitation problem and noisy medical images [12]. The model was based on LSTM, MEVF, and attention mechanisms for text, vision, and fusion phases. They also used a Convolutional Denoising Auto-Encoder (CDAE) to reduce the noise. Their model achieved accuracy figures of 43.9 and 71.5 for open-end and closed-end questions on the VQA-RAD dataset. The model showed its limitations because the text embedding is based on GloVe, which may have a problem with a larger corpus, and they trimmed questions to 12 words.

Most medical VQA utilized transfer learning techniques in the vision phase to avoid data limitations, but the pre-trained models trained on general images differ from medical images. Medical images are also noisy image labels. To solve these problems, Do et al. designed a Multiple Meta-Model Quantifying (MMQ) method for medical VQA [56].

They utilized Model Agonistic Meta-Learning (MAML) to increase meta-data based on auto-annotations. MMQ has three models: meta training for image feature extraction based on MAML; data refinement based on auto-annotation to increase the data and exceed the noisy labels limitation; and meta quantifying that has the decision of meta-model selection for best performance guarantee. They evaluated their model on two datasets: RAD-VQA and PathVQA. The Glove pre-trained model was used for the text phase, whereas SAN and BAN were applied for attention alternately. The model with BAN attention obtained the highest accuracy scores, 48.8% and 67%, on PathVQA and RAD-VQA, respectively.

VQA requires high reliability and performance, which a reliable dataset may achieve. There is only one dataset (VQA-RAD) that experts validated, whereas others were generated by semi-auto-creation or auto-creation [32]. Since VQA-RAD is a small dataset, Liu et al. created a new medical dataset, SLAKE, that was created by specialists and is more extensive than VQA-RAD [32]. Additionally, external knowledge can enhance the performance and robustness of the VQA models. Therefore, they created a medical knowledge base extracted from Wikipedia's large-scale knowledge base. They built two models based on LSTM, VGG, and SAN for the text, vision, and fusion phases. The primary difference between those two models is that one model is based on applying FCN segmentation for images before using VGG. This last method enhanced the accuracy from 72.73% to 75.36% for the English language. They trained the model without FCN segmentation for the Chinese language and achieved 74.27% accuracy. The accuracy levels are still not at the required standard for practical medical applications [32].

Liao et al. and Gong et al., the winners of the last two ImageCLEF challenges in 2020 and 2021, utilized the ensemble methods [55], [57]. Another team in the ImageCLEF Challenge 2021 achieved third place using the ensemble technique with accuracy and BLEU figures of 0.348 and 0.391, respectively. Eslami et al. [79] model was based on converting the VQA problem to an image classification problem and ignoring the text part. It was also the winner of this challenge [57].

E. MODELS WITH V+L PRE-TRAINED MODELS

According to Li et al. [147], V+L pre-trained models in vision and text tasks outperform RNN-CNN models. They compared four V+L pre-trained multi-modals: UNITER [106], LXMERT [107], VisualBERT [92], and PixelBERT [114]. All these pre-trained models used BERT on their text embedding. Since the clinicalBERT pre-trained model has a better effect in the medical field than BERT, Lie et al. supposed that replacing the BERT pre-trained model in the UNITER, LXIMERT, VisualBERT, and PixelBERT pre-trained multi-models with clinicalBERT leading to better multi-model than the original multi-models. However, all new multi-models behaved worse than the original multi-

models. This situation occurred because for various reasons: 1) they trained the models for only 12 epochs, which may not be enough for adjusting the weights; 2) UNITER and PixelBERT have two versions of models: basic and deeper. They configured only the basic ones, which are not the best choice; and 3) in PixelBERT, they froze the vision part and copied the original weights to use them with clinicalBERT text embedding. These weights required adjustment. In the baseline performances of those four V+L pre-trained models, PixelBERT achieved the highest performance in the COCO dataset, whereas visualBERT achieved the worst performance in the same dataset. In Li et al.'s study, PixelBERT achieved the lowest performance in the MIMIC-CXR dataset, whereas visualBERT achieved the highest performance in the same dataset. We believe that this result happened as the reason for the previous limitations in their experiments and choosing the shallower models instead of deeper ones.

F. MODELS BASED ON KNOWLEDGE-BASE

Medical VQA needs information about the diseases and patients' histories. This information is not included in the existing datasets. Therefore, an external knowledge base is needed to enhance the robustness of the models and make them more practical. Lui et al. [32] and Zheng et al. [178] utilized the external knowledge in training their models. Lui et al. [32] created the SLAKE dataset and presented the baseline scores in the created dataset. Zheng et al. [178] designed a model that utilized BERT, VGG-16 with GAP, and BLOCK for the text, vision, and fusion phases, respectively. The model achieved the highest score on the VQA-Med 2019 dataset: 91.2%m 93.8%, 95.7%, 95.9%, and 95.8% for BLEU, accuracy, precision, recall, and F-means. Although the score was high, the dataset on which the model trained had a high bias; no clarification was mentioned in relation to this problem or the use of data visualization to check whether the model learned the alignment between text and vision or if the result stemmed from the dataset bias.

Kovaleva et al. [9] used the patient history trained model, which utilized LSTM and DenseNet-121 for the text and vision parts, on the RedVisdial dataset. They designed three models with the same text and vision parts and three different fusion methods, namely, SAN, LFM, and RVA attentions, that achieved a macro-average score of 34, 33, and 33, respectively. The used patient history was based on only one sentence.

G. VQA WITH OTHER TASKS

Recently, Cong et al. [182] have utilized an image caption task to give summarized information about the image in the medical VQA. This information is embedded and merged with question and image features to enhance the classification task performance. They utilized ResNet-152, BERT, and Progressive Compact Bilinear Interactions (PCBI)for vision featurization, text featurization, and fusion phases, respectively. The method was validated on RAD-VQA and SLAKE datasets, where achieved 69.8 (+8.7), 79.8 (−0.6), 75.8 (+3.1), 80.2 (−1.0), 86.1 (+2.7), and 82.5 (+0.4) for open, close, and overall datasets, respectively. One limitation of this method is that the image caption task affects the wrong information obtained by the final classification.

Tables 16-14 show the the models' comparison based on the the datasets each model fine-tuned on.

VIII. DISCUSSION AND ANALYSIS

Most researchers tend to exploit CNNs for vision feature extraction due to the effectiveness of using deep learning in medical object detection and classification. Some researchers have built CNNs from scratch, whereas others have taken advantage of pre-trained models, such as ResNet, VGGNet, and BERT. However, pre-trained models focused on large image or text datasets weaken model generalizability. Li et al. [147] demonstrated that the VQA model based on the V+L pre-trained model outperforms models based on CNN-RNN models. Thus, the overall performance of the existing models needs to be enhanced. While Vu et al. [7] achieved more than 90% macro accuracy, they achieved a recall of less than 10% on BACH and VQA-Med v2 due to unbalanced data. Zheng et al.'s models achieved an accuracy of 93.8% on VQA-Med v2 [178], but a major limitation with the VQA-Med v2 is a bias. For example, Zhou et al. failed to show whether this result was attained as a result of a dataset bias, which is all the more concerning especially given that Lubna et al. [186] achieved 80.8% accuracy without using any questions whatsoever.

Following a state-of-the-art example, we concluded that factors capable of improving such models involve utilizing ensemble learning in the visual aspect, GAP, and using an external knowledge-base, such as in [178]. We conducted two types of analysis: statistical analysis and SWOT analysis. The statistical analysis was designed to help researchers learn which methods are primarily utilized in a medical VQA, as well as which has a significant impact on improving performance. This analysis could thus help in decision-making processes regarding the techniques researchers require in their models. The SWOT analysis, meanwhile, can help to explain the strengths, weaknesses, opportunities, and threats of a medical VQA, providing insights that can aid future research.

A. STATISTICAL ANALYSIS

In this subsection, we aim to show the statistical analysis of VQA models in the medical field and compare the findings with those from a general field to examine how the latter inspired researchers in a medical context. A statistical analysis of the medical VQA benchmarks is also presented in this discussion.

1) MEDICAL VQA BENCHMARKS

According to the state-of-the-art models presented in Tables 6-14, we detected the frequencies of using each dataset, as shown in Figure 10. Table 15 shows the benchmarks, as well as the best performance achieved in each instance.

TABLE 15. The best performance of the models trained on each dataset.

Dataset	Research	year	BLEU	WBSS	Accuracy	F-score	Recall	Precision
RAD-VQA [10]	Cong et al. [182]	2022				75.8 ±3.1		
VQA-Med V1 [3]	Peng et al. [93]	2018	16.2	18.5				
VQA-Med V2 [31]	Zheng et al.	2020	91.2		93.8	95.8	95.9	95.7
VQA-Med V3 [35]	AIML	2020	54.2		49.6			
VQA-Med V4 [165]	Gong et al. [57]	2021	41.6		38.2			
PathVQA [8]	Do et. al [56]	2021			48.8			
RadVisDial [9]	Kovaleva et al. [9]	2020				34.0		
BACH [7]	Vu et al. [7]	2020			91.42		55.4	71.27
Tools [7]	Vu et al. [7]	2020			96.35		37.85	53.72
IDRiD [7]	Vu et al. [7]	2020			90.3		65.7	53.72
SLAKE [32]	Cong et al. [182]	2022			82.5 (+0.4)			
OVQA [37]	Huang et al. [37]	2022			68.5			
DME [28]	Al-Hadhrami et al. [192]	2023			87.41			
EndoVis-18-VQA [38]	Seenivasa et al. [194]	2023			68.11	46.49	49.69	
Cholec80-VQA [194]	Seenivasan et al. [194]	2023			94.29	74.39	73.39	
P-VQA [43]	Huang et al. [43]	2023			95.45 ±0.13	97.40 ±0.23	97.49 ±0.26	97.43 ±0.24

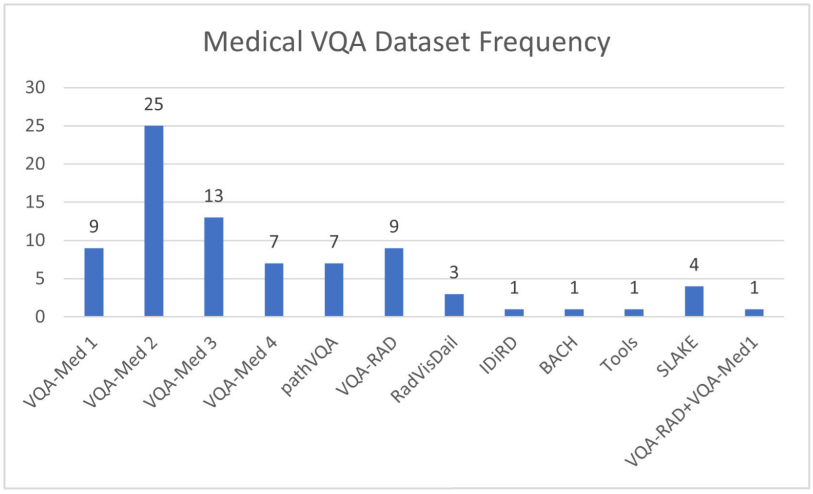
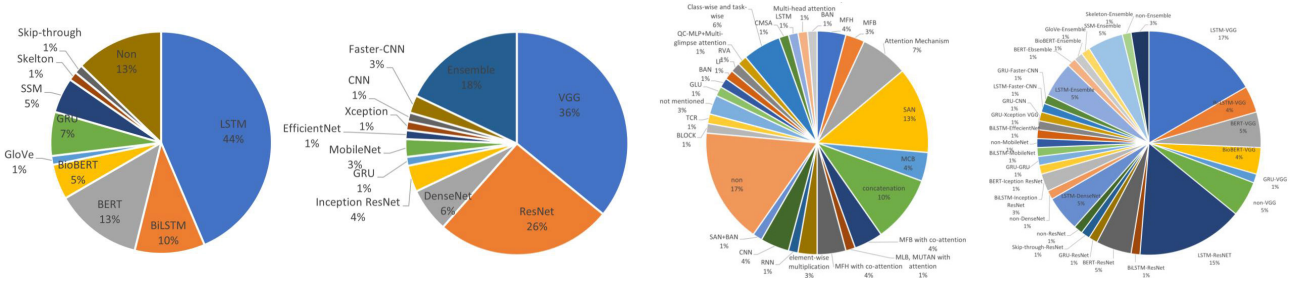


FIGURE 10. Medical VQA dataset frequency.



(a) Text Featurization (b) Vision Featurization (c) Fusion Featurization (d) Text-Vision Featurization

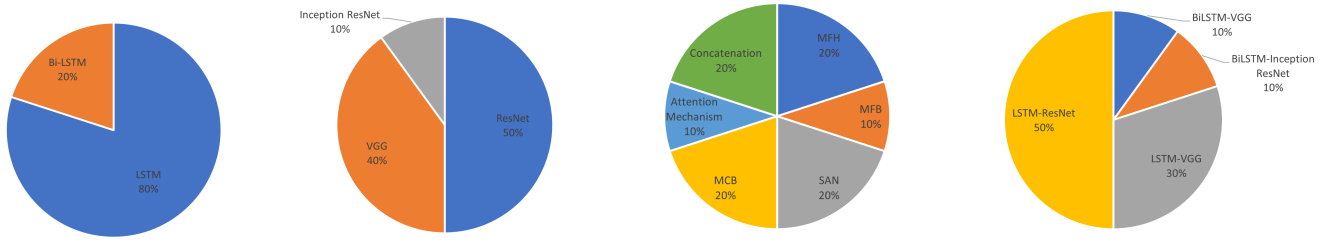
FIGURE 11. Medical VQA dataset frequency.

2) MEDICAL VQA MULTI-MODAL

In Tables 6-14, we compared more than 75 VQA models in terms of structure and performance in the medical field. The findings are summarized in Table 16, which details the vision and text encoding methods frequently used by researchers and highlights which combination is predominantly used. Long short-term memory (LSTM) was the text encoding method most frequently used in medical VQA with a rate

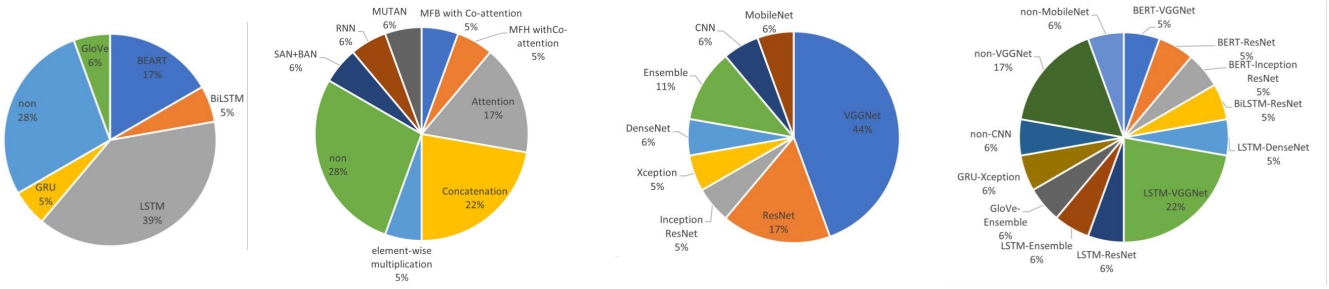
TABLE 16. Text and vision techniques distribution.

	LSTM	Bi-LSTM	BERT	Bio-BERT	GloVe	GRU	SSM	Skelton	Skip-through	Non	Total
VGG	13	3	4	3		1				4	28
ResNet	12	1	4			1			1	1	20
DenseNet	4									1	5
Inception ResNet		2	1								3
GRU						1					1
MobileNet		1								1	2
EfficientNet		1									1
Xception						1					1
CNN										1	1
Faster-CNN	1					1					2
Ensemble	4		1	1	1		4	1		2	11
Total	34	8	10	4	1	5	4	1	1	10	78



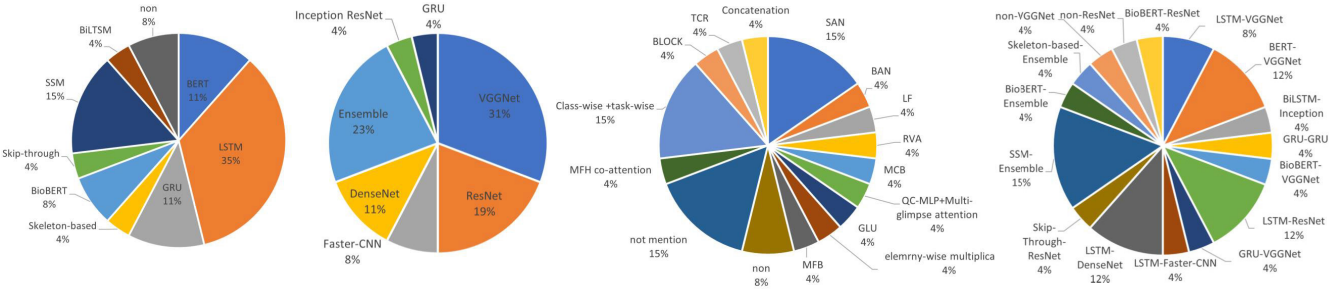
(a) Text Featueization-2018 (b) vision VQA -2018 (c) Fusion VQA-2018 (d) Text and Vision Combination - 2018

FIGURE 12. Medical VQA techniques - 2018 distribution.



(a) Text Featurization (b) Vision Featurization (c) Fusion Featurization (d) Text-Vision Featurization

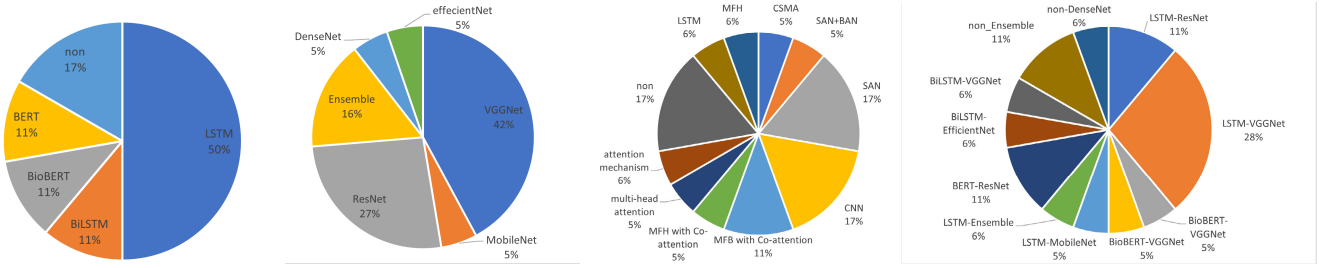
FIGURE 13. Medical VQA techniques-2019 distribution.



(a) Text Featurization (b) Vision Featurization (c) Fusion Featurization (d) Text-Vision Featurization

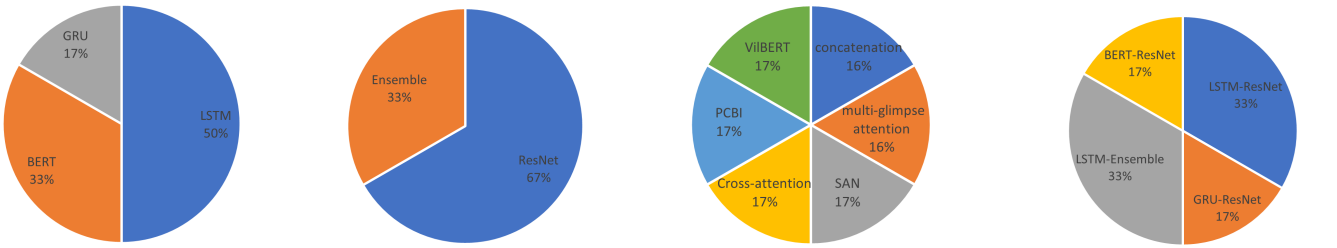
FIGURE 14. Medical VQA techniques-2020 distribution.

of 44%, followed by no text encoding and BERT. The VQA is then converted to the image classification problem with a rate of 13%. Ignoring the text featurization phase in the VQA limits the answer generation, causing a generalization problem within the model. Although some models using this method achieved a significant performance, such as



(a) Text Featueization-2021 (b) vision VQA -2021 (c) Fusion VQA-2021 (d) Text and Vision Combination -2021

FIGURE 15. Medical VQA techniques - 2021 distribution.



(a) Text Featurization Methods 2022-2023 (b) Vision Featurization Methods 2022-2023 (c) Fusion Methods 2022-2023 (d) Text and Vision Combination Methods 2022-2023

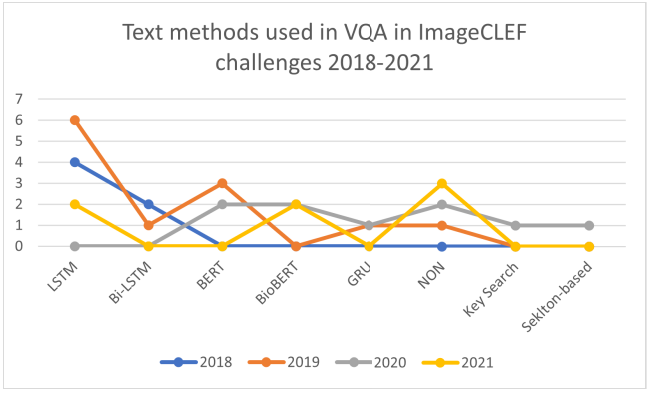
FIGURE 16. Medical VQA techniques - 2022-2023 distribution.

Gong et al. [57], who won the ImageCLEF 2021 challenge, or Al-Sadi et al., who achieved second place in the ImageCLEF 2020 challenge, this does not mean it represents a good choice for a practical VQA, especially in a medical context. Achieving a high score in VQA while ignoring one part of the multi-modal denotes the bias problem in the dataset. Putting aside those two methods, Bi-LSTM and GRU were used, achieving rates of 10% and 7%, respectively, whereas SSM and BioBERT each recorded a rate of 5%. Other methods, such as skeleton-based, skip-through, and Glove, were rarely used. The winning model in ImageCLEF 2020 utilized a skeleton-based method [156].

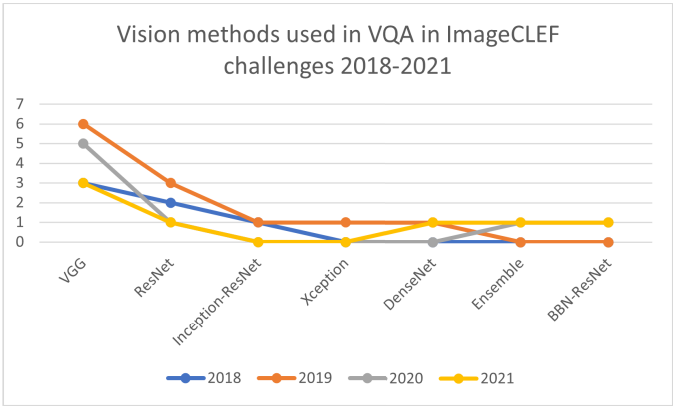
In terms of vision, 11 methods were detected. The VGGNet, ResNet, Ensemble, and DenseNet were the most utilized, with a rate of 95%. The VGGNet was used in 28 out of 78 models, with a rate of 34%. ResNet, Ensemble, and DenseNet were used in 20, 14, and 5 out of 78 models, respectively. While the researchers described a variety of reasons for their choices regarding the multi-modal parts, the performance of all the proposed models in this review shows that these explanations reveal a failure of understanding in terms of how the data are manipulated inside the model. Consequently, a more successful representation explaining both the model's behavior and data visualization inside the model is required. Figure 11 shows the distributions rates of the text featurization, vision, fusion, and text-vision combination methods.

To assess the change in using multi-modal parts over time, we analyzed these methods yearly from 2018 to 2023. Figures 14-18 show the distributions of text, vision, fusion, and text and vision techniques in each instance. Each year, the distributions were 10, 18, 26, 18, 5, and 1 models, respectively. Since there is only one model in 2023, we merge its analysis with the previous year, 2022. We found that VGGNet was the vision pre-trained model primarily used in 2019, 2020, and 2021, with rates of 40%, 35%, and 42%, respectively. By contrast, the ResNet was used in 50% and 67% of the 2018 and 2022-2023 models, while VGGNet was used in 40% in 2018 and not used in 2022-2023. On the other hand, LSTM was the text featurization technique most widely used, with 80%, 39%, 35%, 50%, and 50% of models utilizing the LSTM machine learning method for the text encoding across the six years, respectively. Therefore, the LSTM-CNN combination technique is the most-used multi-modal VQA in the medical field, with rates of 80%, 49%, 36%, 50%, and 33%, respectively, from 2018-2023. However, this high utilization of LSTM-CNN failed to achieve the highest performance on any medical VQA datasets for all four years. The fusion part is the most critical phase in the VQA multi-modal. Different techniques were utilized, as shown in Figures 12-16(c).

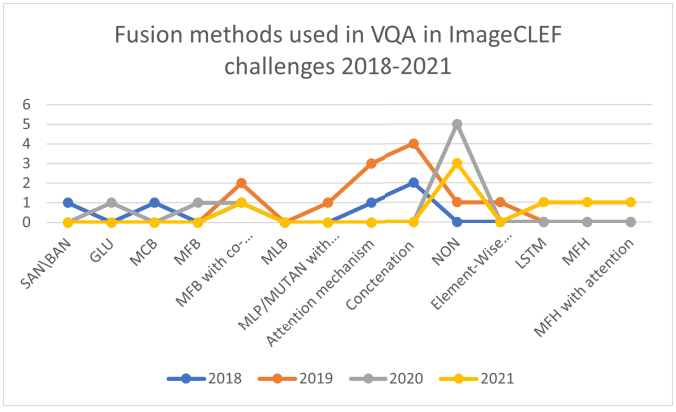
We conducted a statistical analysis for the models used in the ImageCLEF challenges 2018-2021. in 2022, ImageCLEF did not call for VQA challenge. Figure 17 shows the statistical vision, text, and fusion techniques used in these instances based on the models published by their teams. The number of published papers are six [46], [46], [73], [74], [75], [93], twelve [62], [103], [155], [156], [157], [158], [159], [160], [161], [162], [163], [164], eight [35], [61], [166], [167], [168], [169], [170], [171], and seven [57], [79], [80], [165], [172], [173], [174], [175], [176] for 2018, 2019, 2020, and 2021, respectively. In the vision phase, it can be clearly seen



(a) Text methods used in VQA in ImageCLEF challenges 2018-2021

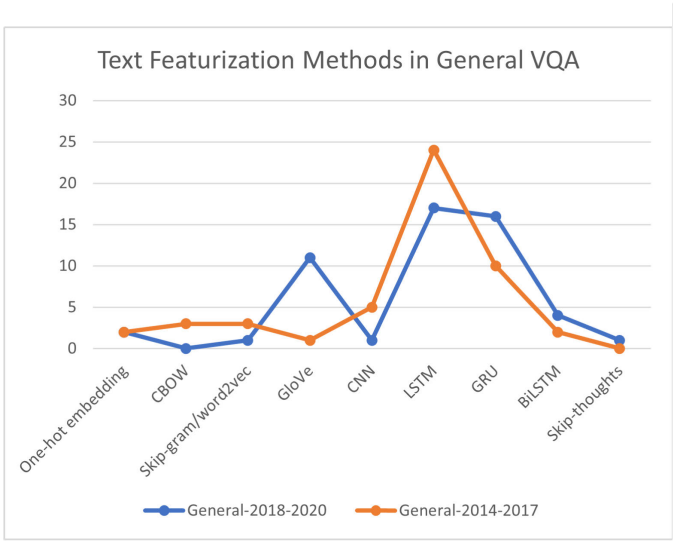


(b) /Vision methods used in VQA in ImageCLEF challenges 2018-2021

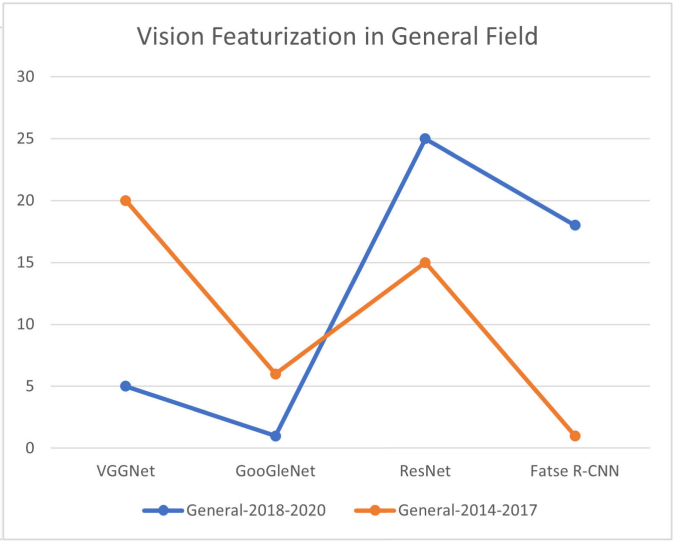


(c) Fusion methods used in VQA in ImageCLEF challenges 2018-2021

FIGURE 17. **ImageCLEF medical VQA 2018-2021 analysis.**



(a) Text Featurization Methods in General VQA



(b) Vision Featurization in general field

FIGURE 18. **General VQA 2014-2020 analysis.**

from Figure 17 that VGG and ResNet pre-trained models were mostly used across all four challenges. Ensemble and BBN-ResNet were used in the two most recent years due to their positive effect on general VQA and more recent medical VQA. The winning models for those two years were both ensemble models.

In terms of the text phase, most participants used LSTM in the 2018 and 2019 challenges before preferring to use transformers like BERT and BioBERT in 2020. However, in 2021, some participants ignored the text part and converted the problem into image classification due to the dataset only containing abnormality questions. The skeleton-based text featurization proved its effectiveness in the 2020 challenge compared to other methods.

Most of the fusion techniques deployed across the challenges involved using an attention mechanism, as shown in Figure 17-(c). In the 2021 challenges, no fusion method was used in most models because converting the problem into images classification or because the authors did not mention it in their papers.

Besides VQA analysis in the medical field, we also conducted an analysis of VQA in the general field for vision and text featurization, based on the most comprehensive survey by Sahani et al. (2021) [23]. Sahani et al. [23] reviewed VQA in the general domain from 2014-2020. Since medical VQA began in 2018, we analyzed the methods into two time periods: 2014-2017 [58], [59], [60], [82], [85], [86], [87], [95], [97], [195], [196], [197], [198], [199], [200], [201], [202], [203], [204], [205], [206], [207], [208], [209], [210], [211], [212], [213], [214], [215], [216], [217], [218], [219] and 2018-2020 [4], [81], [88], [89], [96], [98], [137], [220], [221], [222], [223], [224], [225], [226], [227], [228], [229], [230], [231], [232], [233], [234], [235], [236], [237], [238], [239], [240], [241], [242], [243], [244], [245], [246], [247], [248], [249], [250]. Figure 20 shows the distributions of the textual and visual featurization methods in the general field between 2014-2020. We also compared the text and vision featurization in general if the multi-modal structure followed Sahani et al. [23] to show whether the methods utilized in the general field influenced the researchers in medical VQA.

From Figure 18, it is clear that in terms of the text featurization, LSTM was the most widely used technique from 2014-2017, with a big difference rate with other methods. However, its use was then reduced by approximately 30% from 2018-2020. But this was not the case in the medical field, where 42% of models utilized LSTM as shown in Figure 11. Researchers preferred using the GRU from 2018-2020 instead. However, the GRU was rarely used in medical VQA, with a rate of 6% from the proposed methods included in this review. On the other hand, general VQA models used VGGNet as the vision featurization method from 2014-2017, but from 2018-2020, the general VQA reduced using VGGNet by approximately 75%, increasing the use of ResNet by approximately 66.7% compared to previously. This analysis shows that researchers in the medical field did not follow the general development of VQA.

B. SWOT ANALYSIS

This section presents a SWOT analysis for medical VQA datasets to show their main characteristics, helping researchers to create new datasets or find methods that can overcome their existing limitations. It also presents a SWOT analysis of the existing multi-modal VQA in the medical field, as well as vision and language pre-trained models.

1) MEDICAL VQA DATASET

VQA requires a vast dataset. Simple data with no complex questions will lead to a simple model that cannot answer complex problems. In medical VQA, while recent data is made available annually by ImageCLEF-Med, allowing researchers to either enhance existing data or generate new data, this data remains insufficient in terms of developing a robust and practical model used in the real world [8]. The limitation comes down to the size of the detail, which needs to be sufficiently large to handle various questions. Furthermore, limited data with insufficient information regarding the images or patient history limits the real-world medical VQA-agent system [8]. Although Kovaleva et al. [9] proposed patient history data, this was extracted the based on only one sentence. Furthermore, unbalanced data or biased data are two other data limitations, while the automatic data generation methods used fail to create robust data [10]. These various data barriers sufficiently affect the performance of VQA models. While researchers have proposed different solutions and multi-models to exceed existing borders and enhance overall performance, these efforts remain ineffective.

Table 17 shows a summary of the medical VQA dataset SWOT analysis.

2) MEDICAL VQA SWOT ANALYSIS

Medical VQA is a new field that requires comprehensive analysis in order to achieve a practical VQA-agent that can be trusted by medical staff. Although it represents a new area of research, existing studies have made significant progress, as demonstrated by enhanced performance over the last four years. Various techniques have been developed for the fusion phase of VQA. However, in terms of medical VQA, various weaknesses remain, and the level of performance in the field is still considered too poor to be deployed in real-world settings. A great deal of work must be done to improve performance, especially given the insufficient explanations provided thus far for the errors and model behavior proposed. The leaking of vast trusted datasets that specialists validate also remains an ongoing concern. No V+L pre-trained model has been designed specifically for the medical field, opening opportunities for researchers to find solutions.

The metrics used for VQA are ineffective in terms of open questions, so developing a new metric especially for those question types, is in demand. One limitation related to VQA in the medical field concerns having a large, manually validated dataset, which is a key requirement for creating a trusted medical VQA-agent. Table 18 and Table 19 show the SWOT analysis summary of the V+L pre-trained models and the medical VQA, respectively.

C. CHALLENGES AND RECOMMENDATIONS

According to the SWOT analysis detailed in the previous section, we propose a number of challenges facing VQA

TABLE 17. SWOT analysis on medical VQA dataset.

Strength	Weakness
-publicly large datasets -easiness of creating a new dataset based on the existing medical image dataset. Even automatic generated datasets suffer from errors or bias; it is a solution for difficulty getting large datasets based on specialists.	-leak of an existing dataset with complete information needed in the medical field. -automatic dataset generation leads as a reason for original dataset error. -A real-world medical VQA agent system requires a vast dataset. -lack of diversity and robustness in question-answer pair as a reason for automatic generation dataset. -Dataset bias. -non-equivalent classes dataset
Opportunities	**Threats**
-The lack of an existing dataset with complete information needed in the medical field encourages researchers to develop more powerful techniques of automatic dataset generation. -multi-modality dataset encourages the researchers to design more powerful medical VQA multi-modals.	-limited data with insufficient information about the images or patient history limit the real-world medical VQA-agent system. -automatic dataset generation may lead to errors in the generated dataset. -Difficulty of getting large dataset that checked by specialists. -Simple data with no complex questions leads to a simple model that cannot answer complex problems in the real world. -vast dataset consumes high computations.

TABLE 18. SWOT analysis on medical VQA.

Strength	Weakness
- medical assistant. - the medical VQA-Agent system to help patients understand their X-rays, CT scans, or MRI images. - The medical VQA helps students in the medical field by asking about radiology images and testing themselves. - Accurate models assist doctors and specialists in better understanding the information in an image by asking questions about ambiguous objects that may be present. - V+L pre-trained models handle the problem with small datasets.	- The models are still not accurate enough to be in a real-world VQA-agent system. - No enough explanations of the errors and model behavior. - Leak of vast trusted datasets that specialists validate. - No V+L pre-trained model is designed especially for the medical field. -No specific VQA evaluation metric. - Multi-images questions have not been compromised yet.
Opportunities	**Threats**
- Specialists require precision information. This requirement encourages the researchers to design a medical VQA-agent in the real world, which can give more details and information for specialists. - The lack of specific performance metrics encourages the researchers to create and develop suitable metrics. - Low field performance opens opportunities for researchers to analyze the corresponding errors and enhance the performance.	- The model that is trained on a huge dataset needs high computations. - Misunderstanding of the errors and model behavior may cause a big dilemma in the medical field and make the VQA-Agent not trusted. - Multi-modality datasets add more complexity to the model.

TABLE 19. SWOT analysis on VQA V+L pre-trained models.

Strength	Weakness
- Many V+L pre-trained models are available. Existed model contribute to performance enhancement. - They consider the relation between text and image. - Solve data size limitation.	-The performance still needs to be an enhancement. - It works as a black box that does not allow us to understand how the alignment between text and image has happened. - The errors occurred by the models need explanation. - By experiments, no grantee of performance enhancement over other multi-modal structures.
Opportunities	**Threats**
- the good significant impact of those models encourages research to develop new medical V+L pre-trained models. - encourage the research to interpret the pre-trained model behavior. - and create	- Ambiguity of V+L pre-trained models behaviours. - Selecting the suitable hyper-parameters is still open problem. -

in the medical field and provide recommendations to future researchers.

1) LIMITED DATASETS SIZE

This limitation requires researchers to expand the dataset, either by using other existing datasets or creating a new dataset. Automatic dataset generation helps in creating a vast dataset, but this method has drawbacks. The related limitations are a leak of specialists who validate the new or expanding dataset, original dataset errors, and bias. Another solution regarding the limited dataset size is dataset augmentation classical methods or Generative Adversarial Networks (GANs). We suggest a dataset augmentation method that has not been applied yet. 1this method is based on questions and answers can be paraphrased to augment the dataset. However, even if this approach expands the dataset, it still

may lead to a bias barrier. These challenges mean researchers must spend considerable efforts on creating solutions to the medical dataset generation problem. Transfer learning is one potential solution that may help in solving limited dataset size dilemmas. According to state-of-the-art examples, only one study has utilized a V+L pre-trained model. We recommend using V+L pre-trained models because they are pre-trained on vast V+L datasets and the text is aligned with images.

2) QUESTION DIVERSITY

VQA-RAD [10] and SLAKE [32] have the most diversity among all medical VQA datasets, as shown in Table 4, whereas others suffer from diversity limitations. Future researchers should increase the diversity of the question when creating a new dataset, as insufficient diversity will lead to impractical medical VQA-agents. Increasing the question diversity of the dataset can be achieved via an external source, such as knowledge bases, books, another dataset, or patients' histories. Augmentation using negative or question combinations of more than one question using logical conductive rules increases the dataset diversity. The new dataset, which has more questions diversity than those in the original dataset, needs to be validated by experts. Besides, combining more than one dataset with different question types will increase the diversity.

3) UNIMODALITY BIAS PROBLEM

This limitation denotes the ability to avoid one modality of the multi-modality with significant performance, such as Lubna et al. [186], who achieved 84.6% on VQA-Med 2019 without using the text part in pre-training the model. This limitation can reduce the model's robustness. The bias in text modality is called text prior. Creating a new dataset, or expanding an existing dataset to remove the bias, is one solution to this problem. However, since creating a vast medical VQA benchmark is difficult, researchers need to find methods that contribute to robust the model while reducing bias sensitivity. The ensemble method has been shown to reduce the dataset bias [57], while a variety of methods have been proposed in the general VQA field to aid bias reduction [251], [252], [253], [254], [255]. Yuan's survey has proposed language bias in VQA [256].

4) MULTI-MODALITY DATASET

Most existing medical VQA datasets are multi-modality datasets containing different image formats, i.e., MRI, X-ray, and CT. This multi-modality increases learning difficulty. Splitting the dataset into several single modality datasets and training each one with a different model configuration may be an effective method. However, this approach may exacerbate barriers regarding the limited size of the dataset due to its existence in the original dataset. To solve this limitation, see subsection VIII-A.

5) EXTERNAL KNOWLEDGE

Disease diagnosis may depend on patient history, lab tests, and other information about both the disease itself and its intersection with other diseases. As all existing datasets do not have this information, it would be better to use external knowledge to improve the learning capabilities of future models. Zhou et al. [109] and Liu et al. [32] have built models based on the external knowledge-base. We recommend that the researchers use external knowledge and multiple resources to make the model more practical.

6) MULTIPLE IMAGE

One necessary procedure used in medicine is to follow up on patients' progress by periodically checking their radiology images. As no model has been developed thus far for this, we recommend creating a dataset with multiple images and developing a robust and practical modal to follow up on patient progress.

7) NON-EQUIVALENT CLASSES

The dataset contains non-equivalent classes, meaning that not all classes present in the training set are included in the testing set. This case poses challenges in effectively evaluating the model's performance. For instance, in the case of VQA-RAD and SLAKE, which are predominantly utilized in med-VQA, the training and testing sets consist of 473 and 121 classes and 221 and 33 classes, respectively. This variance of the class number between testing and training sets can not give a trust model performance comparisons because one model may predict the classes are not in the testing set better than those in the testing set, and it is considered poor performance compared to another model that can not detect classes that are not in the testing set. On the other hand, by increasing the examples in the testing set with new classes, the first model, which is considered poorer than the second model, outperforms the second one. Therefore, we encourage researchers to generate, extend, and utilize equivalent class datasets.

8) INTEGRITY WITH AN APPLICABLE MEDICAL APPLICATION

The primary purpose of a medical VQA is to build practical AI-Agents that support the specialist in their decision-making regarding diseases, medical students in their studies, and patients in interpreting their radiology images without the need of specialists. These aims cannot be realized until a robust model with a meager error rate and high generalization has been developed. Researchers must consider all previous comments and threats, as mentioned in Table 9-11.

9) MODEL INTERPRETATION

Model interpretation represents a major challenge for researchers. One obstacle is that there is no explicit model behavior, nor any explanation regarding the reasons for selecting a specific answer. As the model is manipulated as a black box, most existing models have been designed based

TABLE 20. Abbreviation and acronyms.

Beginning of Table	
Abbreviation	Acronyms
AI	artificial intelligence
BERT	Bidirectional Encoder Representations from Transformers
BLUE	BiLingual Evaluation Understudy
CBOW	continuous bag-of-words
CBSS	Concept-based Semantic Similarity
CMF	crossmodal multistep fusion
CT	Copmuted Tomography
DME	Diabetic Macular Edema
DPPNs	dynamic parameter prediction networks
GA	guided attention
GANs	Generative Adversarial Networks
Glove	global vectors
GLU	gating mechanism
G-MLB	Global Multimodal Low-rank Bilinear
G-MLP	Global Multimodal Low-rank Bilinear
GU	gated recurrent units
HOG	histogram of oriented gradients
KR	knowledge representation and reasoning
LF	Late Fusion Network
LSTM	long-sort-term memory
MAML	Model Agonistic Meta-Learning
MCAN	Modular Co-Attention Network
MCAN	Modular Co-Attention
MCAoAN	Modular Co-Attention on Attention Network
MCB	multimodal compact bilinear pooling
METEOR	Metric for Evaluation of Translation with Explicit ORdering
MFB	Multi-modal Factorized Bilinear
MFH	multi-modal factorized highorder pooling
MLP	Multi-modal Low-rank Bilinear
MLP	multi-layer perceptron
MMQ	a Multiple Meta-Model Quantifying
MPT	Mean-Per-Type
MRI	Magnetic resonance imaging
MRNs	Multimodal Residual Network
MUTAN	Multimodal Tucker Fusion for Visual Question Answering
NLP	Natural Language Processing
NMNs	neural module networks
QA	Question answering
QAM	question-guided attention map
QA-MLP	QuestionCentric Multi-modal Low-rank Bilinear
QTA	question type-guided attention
RVA	Recursive Visual Attention Network
SA	d self-attention
SIFT	scale-invariant feature transform
SLAKE	semantically-labeled knowledge-enhanced
SSM	Skeleton-based Sentence Mapping
SVD	singular value decomposition
SWOT	Strengths, Weaknesses, Opportunities, and Threats
V+L	vision-and-language
VGGNet	Very Deep Convolutional Networks
VQA	visual question answering
VQG	visual question generation
WBSS	Word-based Semantic Similarity
WRAN	word-to-region attention network
WUPS	Wu-Palmer Similarity

on trial and error to select and configure the multi-modality parts. Vu et al. [7] have drawn attention to focus areas that can help to select or generate answers. Although this approach was correct in some samples, the result was not, and the researchers could not explain the reasons why.

10) EVALUATION

Accuracy is mostly used in VQA. Even though it fails in terms of BLEU when used with short sentences, as is the case in VQA, it is usually used as a performance metric regarding this problem. Manmadhan and Kovoor suggest using Ngram EVAluation (NEVA), which was proposed by Forsbom in 2003, because NEVA is better suited to short sentences than BLEU while being similar in other aspects [1]. Proposing a new metric that achieves the need for VQA evaluation is a demanding proposition [1], [7], [14], [92]. Paraphrased sentences resulting from the answer generation have to be classified as correct answers. Developing a performance metric for VQA therefore remains an open, ongoing area of concern.

IX. CONCLUSION

VQA is a vision and language field that concerns answering natural language questions about a given image, with medical VQA used to answer questions about medical images. This paper has comprehensively reviewed numerous medical VQA models, structures, and datasets, as well as V+L pre-trained models by comparing more than 75 models with their statistical and SWOT analysis. It also statistically analyzed multi-modality parts in general fields, highlighting how VQA researchers in medical contexts are not inspired by general VQA research. The combination of text and vision methods was analyzed. According to the statistical analysis, 42%, 14%, and 12% of the models studied used LSTM, Non-text, and BiLSTM methods for text encoding, respectively. The most used vision methods were VGG, ResNet, and Ensemble, with rates of 40%, 22%, and 16%, respectively. In terms of the fusion phase, no method was used 14% of the time, while SAN and concatenation methods were used at rates of 13%, and 10%, respectively. We found that LSTM-VGGNet and LSTM-ResNet combinations were primarily used in medical VQA, with 18% and 15% rates, respectively. Besides the statistical analysis of medical VQA 2018-2023, a statistical analysis of medical VQA in each separate year was performed, showing that LSTM and VGGNet were the main methods utilized for text and vision, respectively, in every year except 2018, where ResNet was most utilized.

This analysis shows a clear difference in methods used across different VQA domains. Our SWOT analysis can help future researchers to identify both the strengths and weaknesses they need to be aware of in this field, while also detailing how these weaknesses can create opportunities for them to contribute solutions. Based on the SWOT analysis, recommended areas of future research are as follows: limited datasets size, question diversity, unimodal bias problems, multi-modal datasets, external knowledge, multiple images, integrity in terms of practical medical applications, model interpretation, and evaluation.

ACKNOWLEDGMENT

The authors would like to thank King Saud University and the College of Computer and Information Sciences. Additionally, the authors would like to thank the Deanship of Scientific Research at King Saud University for funding and supporting this research through the initiative of DSR Graduate Students Research Support (GSR). This work was supported in part by KSU and the Center for Complex Engineering Systems (jointly between MIT and KACST).

ABBREVIATION

The following abbreviations are used in this manuscript:

REFERENCES

[1] S. Manmadhan and B. C. Kovoor, "Visual question answering: A state-of-the-art review," *Artif. Intell. Rev.*, vol. 53, no. 8, pp. 5705–5745, Apr. 2020.
[2] D. H. Park, L. A. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, and M. Rohrbach, "Multimodal explanations: Justifying decisions and pointing to the evidence," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 8779–8788.
[3] S. A. Hasan, Y. Ling, O. Farri, J. Liu, H. Müller, and M. Lungren, "Overview of ImageCLEF 2018 medical domain visual question answering task," in *Proc. CLEF Working Notes*, Sep. 2018, pp. 1–9.
[4] Y. Xi, Y. Zhang, S. Ding, and S. Wan, "Visual question answering model based on visual relationship detection," *Signal Process., Image Commun.*, vol. 80, Feb. 2020, Art. no. 115648.
[5] X. Wang, Y. Liu, C. Shen, C. C. Ng, C. Luo, L. Jin, C. S. Chan, A. van den Hengel, and L. Wang, "On the general value of evidence, and bilingual scene-text visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2020, pp. 10123–10132.
[6] S.-H. Chou, W.-L. Chao, W.-S. Lai, M. Sun, and M.-H. Yang, "Visual question answering on 360deg images," in *Proc. IEEE/CVF winter Conf. Appl. Comput. Vis.*, Jul. 2020, pp. 1607–1616.
[7] M. H. Vu, T. Löfstedt, T. Nyholm, and R. Sznitman, "A question-centric model for visual question answering in medical imaging," *IEEE Trans. Med. Imag.*, vol. 39, no. 9, pp. 2856–2868, Sep. 2020.
[8] X. He, Y. Zhang, L. Mou, E. Xing, and P. Xie, "PathVQA: 30000+ questions for medical visual question answering," 2020, *arXiv:2003.10286*.
[9] O. Kovaleva, C. Shivade, S. Kashyap, K. Kanjaria, J. Wu, D. Ballah, A. Coy, and A. Karargyris, "Towards visual dialog for radiology," in *Proc. 19th SIGBioMed Workshop Biomed. Lang. Process.* Stroudsburg, PA, USA: Association for Computational Linguistics (ACL), Jul. 2020, pp. 60–69.
[10] J. J. Lau, S. Gayen, A. B. Abacha, and D. Demner-Fushman, "Data descriptor: A dataset of clinically generated visual questions and answers about radiology images," *Sci. Data*, vol. 5, no. 1, pp. 1–10, Nov. 2018.
[11] W. Jifara, F. Jiang, S. Rho, M. Cheng, and S. Liu, "Medical image denoising using convolutional neural network: A residual learning approach," *J. Supercomput.*, vol. 75, no. 2, pp. 704–718, Feb. 2019.
[12] B. D. Nguyen, T.-T. Do, B. X. Nguyen, T. Do, E. Tjiputra, and Q. D. Tran, "Overcoming data limitation in medical visual question answering," in *Proc. MICCAI 22nd Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Shenzhen, China, Cham, Switzerland: Springer, 2019, pp. 522–530.
[13] L.-M. Zhan, B. Liu, L. Fan, J. Chen, and X.-M. Wu, "Medical visual question answering via conditional reasoning," in *Proc. 28th ACM Int. Conf. Multimedia*, Seattle, WA, USA, Oct. 2020, pp. 2345–2354.
[14] F. Ren and Y. Zhou, "CGMVQA: A new classification and generative model for medical visual question answering," *IEEE Access*, vol. 8, pp. 50626–50636, 2020.
[15] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. van den Hengel, "Visual question answering: A survey of methods and datasets," *Comput. Vis. Image Understand.*, vol. 163, pp. 21–40, Oct. 2017.

[16] K. Kafle and C. Kanan, "Visual question answering: Datasets, algorithms, and future challenges," *Comput. Vis. Image Understand.*, vol. 163, pp. 3–20, Oct. 2017.

[17] A. K. Gupta, "Survey of visual question answering: Datasets and techniques," 2017, *arXiv:1705.03865*.

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

[19] C. Patil and M. Patwardhan, "Visual question generation: The state of the art," *ACM Comput. Surveys*, vol. 53, no. 3, pp. 1–22, May 2021.

[20] Y. Srivastava, V. Murali, S. R. Dubey, and S. Mukherjee, "Visual question answering using deep learning: A survey and performance analysis," in *Proc. 5th Int. Conf. Comput. Vis. Image Process. (CVIP)*. Prayagraj, India, Singapore: Springer, 2020, pp. 75–86.

[21] Y. Zou and Q. Xie, "A survey on VQA: Datasets and approaches," in *Proc. 2nd Int. Conf. Inf. Technol. Comput. Appl. (ITCA)*, Guangzhou, China, Dec. 2020, pp. 289–297.

[22] H. Sharma and A. S. Jalal, "A survey of methods, datasets and evaluation metrics for visual question answering," *Image Vis. Comput.*, vol. 116, Dec. 2021, Art. no. 104327.

[23] M. Sahani, P. Singh, S. Jangpangi, and S. Kumar, "A survey on representation learning in visual question answering," in *Proc. Int. Conf. Mach. Learn. Big Data Analytics (IMLBDA)*. Cham, Switzerland: Springer, 2021, pp. 326–336.

[24] Z. Lin, D. Zhang, Q. Tao, D. Shi, G. Haffari, Q. Wu, M. He, and Z. Ge, "Medical visual question answering: A survey," 2021, *arXiv:2111.10056*.

[25] M. S. Sunny and W. Katiyar, "A survey on visual questioning answering: Datasets, approaches and models," *Int. J. Sci. Technol. Res.*, vol. 9, no. 2, pp. 3919–3923, Jan. 2020.

[26] S. S. N. Mohamed and K. Srinivasan, "A comprehensive interpretation for medical VQA: Datasets, techniques, and challenges," *J. Intell. Fuzzy Syst.*, vol. 44, no. 4, pp. 5803–5819, Apr. 2023.

[27] Z. Lin, D. Zhang, Q. Tao, D. Shi, G. Haffari, Q. Wu, M. He, and Z. Ge, "Medical visual question answering: A survey," *Artif. Intell. Med.*, vol. 143, Sep. 2023, Art. no. 102611.

[28] S. Tascon-Morales, P. Márquez-Neila, and R. Sznitman, "Consistency-preserving visual question answering in medical imaging," in *Proc. 25th Int. Conf. Med. Image Comput. Comput. Assist. Intervent.–MICCAI*, Singapore, Cham, Switzerland: Springer, 2022, pp. 386–395.

[29] M. Sarrouti, A. Ben Abacha, and D. Demner-Fushman, "Goal-driven visual question generation from radiology images," *Information*, vol. 12, no. 8, p. 334, Aug. 2021.

[30] M. Sarrouti, A. Ben Abacha, and D. Demner-Fushman, "Visual question generation from radiology images," in *Proc. 1st Workshop Adv. Lang. Vis. Res.*, 2020, pp. 12–18.

[31] A. B. Abacha, S. A. Hasan, V. V. Datla, J. Liu, D. Demner-Fushman, and H. Müller, "VQA-med: Overview of the medical visual question answering task at ImageCLEF 2019," in *Proc. CLEF*, Sep. 2019, pp. 1–11.

[32] B. Liu, L.-M. Zhan, L. Xu, L. Ma, Y. Yang, and X.-M. Wu, "Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering," in *Proc. IEEE 18th Int. Symp. Biomed. Imag. (ISBI)*, Nice, France, Apr. 2021, pp. 1650–1654.

[33] H. Gong, G. Chen, S. Liu, Y. Yu, and G. Li, "Cross-modal self-attention with multi-task pre-training for medical visual question answering," 2021, *arXiv:2105.00136*.

[34] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Comput. Vis.–ECCV 13th Eur. Conf.*, Zurich, Switzerland, Cham, Switzerland: Springer, Sep. 2014, pp. 740–755.

[35] A. B. Abacha, V. V. Datla, S. A. Hasan, D. Demner-Fushman, and H. Müller, "Overview of the VQA-med task at ImageCLEF 2020: Visual question answering and generation in the medical domain," in *Proc. CLEF Conf. Labs Eval. Forum*, 2020, pp. 1–9.

[36] P. Porwal, S. Pachade, R. Kamble, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, and F. Meriaudeau, "Indian diabetic retinopathy image dataset (IDRiD): A database for diabetic retinopathy screening research," *Data*, vol. 3, no. 3, p. 25, Jul. 2018.

[37] Y. Huang, X. Wang, F. Liu, and G. Huang, "OVQA: A clinically generated visual question answering dataset," in *Proc. 45th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Madrid, Spain, Jul. 2022, pp. 2924–2938.

[38] L. Seenivasan, M. Islam, A. K. Krishna, and H. Ren, "Surgical-VQA: Visual question answering in surgical scenes using transformer," in *Proc. 25th Int. Conf. Med. Image Comput. Comput. Assist. Intervent.–MICCAI*, Singapore, Cham, Switzerland: Springer, Sep. 2022, pp. 33–43.

[39] E. Decencière, G. Cazuguel, X. Zhang, G. Thibault, J.-C. Klein, F. Meyer, B. Marcotegui, G. Quellec, M. Lamard, R. Danno, D. Elie, P. Massin, Z. Viktor, A. Erginay, B. Laÿ, and A. Chabouis, "TeleOphta: Machine learning and image processing methods for teleophthalmology," *IRBM*, vol. 34, no. 2, pp. 196–203, Apr. 2013.

[40] M. Allan, S. Kondo, S. Bodenstedt, S. Leger, R. Kadkhodamohammadi, I. Luengo, F. Fuentes, E. Flouty, A. Mohammed, and M. Pedersen, "2018 Robotic scene segmentation challenge," 2020, *arXiv:2001.11190*.

[41] M. Islam, L. Seenivasan, L. C. Ming, and H. Ren, "Learning and reasoning with the graph structure representation in robotic surgery," in *Proc. 23rd Int. Conf. Med. Image Comput. Comput. Assist. Intervent.–MICCAI*, Lima, Peru, Cham, Switzerland: Springer, Oct. 2020, pp. 627–636.

[42] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. de Mathelin, and N. Padoy, "EndoNet: A deep architecture for recognition tasks on laparoscopic videos," *IEEE Trans. Med. Imag.*, vol. 36, no. 1, pp. 86–97, Jan. 2017.

[43] J. Huang, Y. Chen, Y. Li, Z. Yang, X. Gong, F. L. Wang, X. Xu, and W. Liu, "Medical knowledge-based network for patient-oriented visual question answering," *Inf. Process. Manage.*, vol. 60, no. 2, Mar. 2023, Art. no. 103241.

[44] Z. Wu and M. Palmer, "Verb semantics and lexical selection," 1994, *arXiv:cmp-lg/9406033*.

[45] G. Soğancıoğlu, H. Öztürk, and A. Özgür, "BIOSSES: A semantic sentence similarity estimation system for the biomedical domain," *Bioinformatics*, vol. 33, no. 14, pp. i49–i58, Jul. 2017.

[46] A. B. Abacha, S. Gayen, J. J. Lau, S. Rajaraman, and D. Demner-Fushman, "NLM at ImageCLEF 2018 visual question answering in the medical domain," in *Proc. Work. Notes CLEF Conf. Labs Eval. Forum*, Lugano, Switzerland, 2018, pp. 1–10.

[47] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, Philadelphia, PA, USA, 2002, pp. 311–318.

[48] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, vol. 2, Aug. 1999, pp. 1150–1157.

[49] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Diego, CA, USA, Jul. 2005, pp. 886–893.

[50] R. Lienhart and J. Maydt, "An extended set of Haar-like features for rapid object detection," in *Proc. Int. Conf. Image Process.*, Sep. 2002, pp. 900–903.

[51] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[52] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[53] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1–9.

[54] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[55] Z. Liao, Q. Wu, C. Shen, A. van den Hengel, and J. Verjans, "AIML at VQA-Med 2020: Knowledge inference via a skeleton-based sentence mapping approach for medical domain visual question answering," in *Proc. Work. Notes CLEF Conf. Labs Eval. Forum*, Thessaloniki, Greece, Sep. 2020, pp. 1–14.

[56] T. Do, B. X. Nguyen, E. Tjiputra, M. Tran, Q. D. Tran, and A. Nguyen, "Multiple meta-model quantifying for medical visual question answering," 2021, *arXiv:2105.08913*.

[57] H. Gong, R. Huang, G. Chen, and G. Li, "SYSU-HCP at VQA-Med 2021: A data-centric model with efficient training methodology for medical visual question answering," *Proc. Work. Notes CLEF Conf. Labs Eval. Forum*, vol. 201, Sep. 2021, pp. 1–11.

[58] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Austin, TX, USA, 2016, pp. 457–468.

[59] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling," in *Proc. 5th Int. Conf. Learn. Represent. (ICLR)*, Oct. 2016, pp. 1–6.
[60] H. Ben-Younes, R. Cadene, M. Cord, and N. Thome, "MUTAN: Multimodal tucker fusion for visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 2631–2639.
[61] H. K. Verma and S. Ramachandran, "Harendrakv at VQA-med 2020: Sequential VQA with attention for medical visual question answering," in *Proc. Work. Notes CLEF Conf. Labs Eval. Forum*, Thessaloniki, Greece, Sep. 2020, pp. 1–7.
[62] R. Bounaama and M. E. A. Abderrahim, "Tlemcen university at ImageCLEF 2019 visual question answering task," in *Proc. Work. Notes CLEF Conf. Labs Eval. Forum*, Lugano, Switzerland, Sep. 2019, pp. 1–6.
[63] G. A. Miller and W. G. Charles, "Contextual correlates of semantic similarity," *Lang. Cognit. Processes*, vol. 6, no. 1, pp. 1–28, Jan. 1991.
[64] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, Sep. 1936.
[65] W. Xu and A. Rudnicky, "Can artificial neural networks learn language models?" in *Proc. 6th Int. Conf. Spoken Lang. Process. (ICSLP)*, Beijing, China, Oct. 2000, pp. 1–4.
[66] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.
[67] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. neural Inf. Process. Syst.*, vol. 26, 2013, pp. 1–6.
[68] Google. (2013). *Word2Vec*. Accessed: Nov. 2023. [Online]. Available: https://code.google.com/archive/p/word2vec/
[69] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[70] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1724–1734.
[71] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Doha, Doha, Qatar, 2014, pp. 1532–1543.
[72] (2020). *Challenge-Pathology Visual Question Answering-Grand Challenge*. Accessed Nov. 2023. [Online]. Available: https://pathvqachallenge.grand-hallenge.org/PathVQA_challenge/
[73] I. Allaouzi, B. Benamrou, M. Benamrou, and M. B. Ahmed, "Deep neural networks and decision tree classifier for visual question answering in the medical domain," in *Proc. CLEF (Working Notes)*, Sep. 2018, pp. 1–7.
[74] Y. Zhou, X. Kang, and F. Ren, "Employing inception-ResNet-v2 and Bi-LSTM for medical domain visual question answering," in *Proc. Work. Notes CLEF Conf. Labs Eval. Forum*, Avignon, France, Sep. 2018, pp. 1–11.
[75] B. Talafha and M. Al-Ayyoub, "Just at VQA-med: A VGG-Seq2Seq model," in *Proc. Work. Notes CLEF Conf. Labs Eval. Forum*, Avignon, France, Sep. 2018, pp. 1–8.
[76] R. Kiros, Y. Zhu, R. Salakhutdinov, R. S. Zemel, A. Torralba, R. Urtasun, and S. Fidler, "Skip-thought vectors," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 3294–3302.
[77] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. 33rd Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2019, pp. 5753–5763.
[78] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[79] S. Eslami, G. de Melo, and C. Meinel, "Teams at VQA-Med 2021: BBN-orchestra for long-tailed medical visual question answering," in *Proc. Work. Notes CLEF Conf. Labs Eval. Forum*, Bucharest, Romania, Sep. 2021, pp. 1211–1217.
[80] R. Schilling, P. Messina, D. Parra, and H. Löbel, "PUC Chile team at VQA-Med 2021: Approaching VQA as a classification task via fine-tuning a pretrained CNN," in *Proc. Work. Notes CLEF Conf. Labs Eval. Forum*, Bucharest, Romania, Sep. 2021, pp. 1346–1351.
[81] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, "Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 5947–5959, Dec. 2018.
[82] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual question answering," in *Proc. ICCV*, Araucano Park, Chile, 2015, pp. 2425–2433.
[83] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A deep learning approach to visual question answering," *Int. J. Comput. Vis.*, vol. 125, nos. 1–3, pp. 110–135, Dec. 2017.
[84] K. Saito, A. Shin, Y. Ushiku, and T. Harada, "DualNet: Domain-invariant network for visual question answering," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Hong Kong, Jul. 2017, pp. 829–834.
[85] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Neural module networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Las Vegas, NV, USA: Caesars Palace, Jun. 2016, pp. 39–48.
[86] H. Noh, P. H. Seo, and B. Han, "Image question answering using convolutional neural network with dynamic parameter prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA: Caesars Palace, Jun. 2016, pp. 30–38.
[87] J.-H. Kim, S.-W. Lee, D. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang, "Multimodal residual learning for visual QA," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–6.
[88] M. Lao, Y. Guo, H. Wang, and X. Zhang, "Cross-modal multistep fusion network with co-attention for visual question answering," *IEEE Access*, vol. 6, pp. 31516–31524, 2018.
[89] Y. Bai, J. Fu, T. Zhao, and T. Mei, "Deep attention neural tensor network for visual question answering," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 20–35.
[90] M. Narasimhan and A. G. Schwing, "Straight to the facts: Learning knowledge base retrieval for factual visual question answering," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 451–468.
[91] L. Chen, X. Yan, J. Xiao, H. Zhang, S. Pu, and Y. Zhuang, "Counterfactual samples synthesizing for robust visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10797–10806.
[92] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "VisualBERT: A simple and performant baseline for vision and language," Aug. 2019, *arXiv:1908.03557*.
[93] Y. Peng, F. Liu, and M. P. Rosen, "UMASS at ImageCLEF medical visual question answering (Med-VQA) 2018 task," in *Proc. Work. Notes CLEF Conf. Labs Eval. Forum*, Avignon, France, Sep. 2018, pp. 1–9.
[94] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 289–297.
[95] K. Chen, J. Wang, L.-C. Chen, H. Gao, W. Xu, and R. Nevatia, "ABC-CNN: An attention based convolutional neural network for visual question answering," 2015, *arXiv:1511.05960*.
[96] Y. Shi, T. Furlanello, S. Zha, and A. Anandkumar, "Question type guided attention in visual question answering," in *Proc. ECCV*, Munich, Germany, Sep. 2018, pp. 151–166.
[97] K. J. Shih, S. Singh, and D. Hoiem, "Where to look: Focus regions for visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Las Vegas, NV, USA: Caesars Palace, Jun. 2016, pp. 4613–4621.
[98] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear attention networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–6.
[99] D.-K. Nguyen and T. Okatani, "Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 6087–6096.
[100] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 6274–6283.
[101] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 4633–4642.
[102] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 933–941.
[103] M. H. Vu, R. Sznitman, T. Nyholm, and T. Löfstedt, "Ensemble of streamlined bilinear visual question answering models for the imageclef 2019 challenge in the medical domain," in *Proc. CLEF Conf. Labs Eval. Forum*, Lugano, Switzerland, vol. 2380, Sep. 2019, pp. 1–6.

[104] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 248–255.

[105] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[106] Y.-C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, Sep. 2020, pp. 104–120.

[107] H. Tan and M. Bansal, "LXMERT: Learning cross-modality encoder representations from transformers," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 5100–5111.

[108] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," 2019, *arXiv:1908.02265*.

[109] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, "Unified vision-language pre-training for image captioning and VQA," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 13041–13049.

[110] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "VL-BERT: Pre-training of generic visual-linguistic representations," 2019, *arXiv:1908.08530*.

[111] J. Guo, C. Zhu, Y. Zhao, H. Wang, Y. Hu, X. He, and D. Cai, "LAMP: Label augmented multimodal pretraining," 2020, *arXiv:2012.04446*.

[112] J. Cho, J. Lu, D. Schwenk, H. Hajishirzi, and A. Kembhavi, "X-LXMERT: Paint, caption and answer questions with multi-modal transformers," 2020, *arXiv:2009.11278*.

[113] Y. Li, Y. Pan, T. Yao, J. Chen, and T. Mei, "Scheduled sampling in vision-language pretraining with decoupled encoder–decoder network," 2021, *arXiv:2101.11562*.

[114] Z. Huang, Z. Zeng, B. Liu, D. Fu, and J. Fu, "Pixel-BERT: Aligning image pixels with text by deep multi-modal transformers," 2020, *arXiv:2004.00849*.

[115] F. Yu, J. Tang, W. Yin, Y. Sun, H. Tian, H. Wu, and H. Wang, "ERNIE-ViL: Knowledge enhanced vision-language representations through scene graph," 2020, *arXiv:2006.16934*.

[116] S. Zhang, T. Jiang, T. Wang, K. Kuang, Z. Zhao, J. Zhu, J. Yu, H. Yang, and F. Wu, "DeVLBert: Learning deconfounded visio-linguistic representations," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 4373–4382.

[117] F. Luo, P. Yang, S. Li, X. Ren, and X. Sun, "CAPT: Contrastive pre-training for learning denoised sequence representations," 2020, *arXiv:2010.06351*.

[118] W. Li, C. Gao, G. Niu, X. Xiao, H. Liu, J. Liu, H. Wu, and H. Wang, "UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning," 2020, *arXiv:2012.15409*.

[119] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao, "VinVL: Revisiting visual representations in vision-language models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Montreal, QC, Canada, Jun. 2021, pp. 5575–5584.

[120] C. Li, M. Yan, H. Xu, F. Luo, W. Wang, B. Bi, and S. Huang, "SemVLP: Vision-language pre-training by aligning semantics at multiple levels," 2021, *arXiv:2103.07829*.

[121] C. Kervadec, G. Antipov, M. Baccouche, and C. Wolf, "Weak supervision helps emergence of word-object alignment and improves vision-language tasks," 2019, *arXiv:1912.03063*.

[122] J. Lin, A. Yang, Y. Zhang, J. Liu, J. Zhou, and H. Yang, "InterBERT: Vision-and-language interaction for multi-modal pretraining," 2020, *arXiv:2003.13198*.

[123] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, and F. Wei, "OSCAR: Object-semantics aligned pre-training for vision-language tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Cham, Switzerland: Springer, 2020, pp. 121–137.

[124] E. Bugliarello, R. Cotterell, N. Okazaki, and D. Elliott, "Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language BERTs," *Trans. Assoc. Comput. Linguistics*, vol. 9, pp. 978–994, Sep. 2021.

[125] W. Kim, B. Son, and I. Kim, "ViLT: Vision-and-language transformer without convolution or region supervision," 2021, *arXiv:2102.03334*.

[126] M. Zhuge, D. Gao, D.-P. Fan, L. Jin, B. Chen, H. Zhou, M. Qiu, and L. Shao, "Kaleido-BERT: Vision-language pre-training on fashion domain," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12642–12652.

[127] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, and J. Clark, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2021, pp. 8748–8763.

[128] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 1–6.

[129] L. Yuan, D. Chen, Y.-L. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, and C. Li, "Florence: A new foundation model for computer vision," 2021, *arXiv:2111.11432*.

[130] C. Alberti, J. Ling, M. Collins, and D. Reitter, "Fusion of detected objects in text for visual question answering," 2019, *arXiv:1908.05054*.

[131] Y. Wang, S. Joty, M. R. Lyu, I. King, C. Xiong, and S. C. H. Hoi, "VD-BERT: A unified vision and dialog transformer with BERT," 2020, *arXiv:2004.13278*.

[132] V. Murahari, D. Batra, D. Parikh, and A. Das, "Large-scale pretraining for visual dialog: A simple state-of-the-art baseline," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, Aug. 2020, pp. 336–352.

[133] W. Hao, C. Li, X. Li, L. Carin, and J. Gao, "Towards learning a generic agent for vision-and-language navigation via pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13134–13143.

[134] D. Qi, L. Su, J. Song, E. Cui, T. Bharti, and A. Sacheti, "ImageBERT: Cross-modal pre-training with large-scale weak-supervised image-text data," 2020, *arXiv:2001.07966*.

[135] Q. Xia, H. Huang, N. Duan, D. Zhang, L. Ji, Z. Sui, E. Cui, T. Bharti, and M. Zhou, "XGPT: Cross-modal generative pre-training for image captioning," in *Proc. CCF Int. Conf. Natural Lang. Process. Chin. Comput.*, Qingdao, China, Cham, Switzerland: Springer, Oct. 2021, pp. 786–797.

[136] T. Scialom, P. Bordes, P.-A. Dray, J. Staiano, and P. Gallinari, "Bert can see out of the box: On the cross-modal transferability of text representations," 2020, *arXiv:2002.10832*.

[137] J. Yu, Z. Zhu, Y. Wang, W. Zhang, Y. Hu, and J. Tan, "Cross-modal knowledge reasoning for knowledge-based visual question answering," *Pattern Recognit.*, vol. 108, Dec. 2020, Art. no. 107563.

[138] H. Singh and S. Shekhar, "STL-CQA: Structure-based transformers with localization and encoding for chart question answering," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 3275–3284.

[139] R. Tanaka, K. Nishida, and S. Yoshida, "VisualMRC: Machine reading comprehension on document images," 2021, *arXiv:2101.11272*.

[140] M.-J. Chiou, R. Zimmermann, and J. Feng, "Visual relationship detection with visual-linguistic knowledge from multimodal representations," *IEEE Access*, vol. 9, pp. 50441–50451, 2021.

[141] J. Lu, V. Goswami, M. Rohrbach, D. Parikh, and S. Lee, "12-in-1: Multi-task vision and language representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10434–10443.

[142] J. Cho, J. Lei, H. Tan, and M. Bansal, "Unifying vision-and-language tasks via text generation," in *Proc. Int. Conf. Mach. Learn.*, Qingdao, China, Oct. 2021, pp. 1931–1942.

[143] C. Ross, B. Katz, and A. Barbu, "Measuring social biases in grounded vision and language embeddings," 2020, *arXiv:2002.08911*.

[144] J. Yang, J. Duan, S. Tran, Y. Xu, S. Chanda, L. Chen, B. Zeng, T. Chilimbi, and J. Huang, "Vision-language pre-training with triple contrastive learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 15650–15659.

[145] T. Wang, W. Jiang, Z. Lu, F. Zheng, R. Cheng, C. Yin, and P. Luo, "VLMixer: Unpaired vision-language pre-training via cross-modal cutmix," in *Proc. Int. Conf. Mach. Learn.*, Baltimore, MD, USA, Jul. 2022, pp. 22680–22690.

[146] F. Liu, X. Wu, S. Ge, X. Ren, W. Fan, X. Sun, and Y. Zou, "DiMBERT: Learning vision-language grounded representations with disentangled multimodal-attention," *ACM Trans. Knowl. Discovery from Data*, vol. 16, no. 1, pp. 1–19, Feb. 2022.

[147] Y. Li, H. Wang, and Y. Luo, "A comparison of pre-trained vision-and-language models for multimodal representation learning across medical images and reports," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2020, pp. 1999–2004.

[148] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. B. A. McDermott, "Publicly available clinical BERT embeddings," 2019, *arXiv:1904.03323*.
[149] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. Int. Conf. Mach. Learn.*, Baltimore, MD, USA, Jul. 2022, pp. 12888–12900.
[150] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang, "OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," in *Int. Conf. Mach. Learn.*, Baltimore, MD, USA, Jul. 2022, pp. 23318–23340.
[151] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "CoCa: Contrastive captioners are image-text foundation models," 2022, *arXiv:2205.01917*.
[152] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, S. Singhal, S. Som, and F. Wei, "Image as a foreign language: BEIT pretraining for vision and vision-language tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 19175–19186.
[153] X. Chen, X. Wang, S. Changpinyo, A. J. Piergiovanni, P. Padlewski, D. Salz, S. Goodman, A. Grycner, B. Mustafa, and L. Beyer, "PaLI: A jointly-scaled multilingual language-image model," 2022, *arXiv:2209.06794*.
[154] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," 2023, *arXiv:2301.12597*.
[155] Y. Zhou, X. Kang, and F. Ren, "TUA1 at ImageCLEF 2019 VQA-Med: A classification and generation model based on transfer learning," in *Proc. Work. Notes CLEF Conf. Labs Eval. Forum*, Lugano, Switzerland, Sep. 2019, pp. 1–7.
[156] X. Yan, L. Li, C. Xie, J. Xiao, and L. Gu, "Zhejiang university at ImageCLEF 2019 visual question answering in the medical domain," in *Proc. Work. Notes CLEF Conf. Labs Eval. Forum*, Lugano, Switzerland, Sep. 2019, pp. 1–6.
[157] A. Thanki and K. Makkithaya, "Mit Manipal at ImageCLEF 2019 visual question answering in medical domain," in *Proc. Work. Notes CLEF Conf. Labs Eval. Forum*, Lugano, Switzerland, Sep. 2019, pp. 1–9.
[158] A. Turner and A. Spanier, "LSTM in VQA-med, is it really needed? JCE study on the ImageCLEF 2019 dataset," in *Proc. Work. Notes CLEF Conf. Labs Eval. Forum*, Lugano, Switzerland, Sep. 2019, pp. 1–9.
[159] M. Bansal, T. Gadgil, R. Shah, and P. Verma, "Medical visual question answering at image clef 2019-VQA med," in *Proc. Work. Notes CLEF Conf. Labs Eval. Forum*, Lugano, Switzerland, Sep. 2019, pp. 1–9.
[160] L. Shi, F. Liu, and M. P. Rosen, "Deep multimodal learning for medical visual question answering," in *Proc. Work. Notes CLEF Conf. Labs Eval. Forum*, Lugano, Switzerland, Sep. 2019, pp. 1–6.
[161] S. Liu, X. Ou, J. Che, X. Zhou, and H. Ding, "An Xception-GRU model for visual question answering in the medical domain," in *Proc. Work. Notes CLEF Conf. Labs Eval. Forum*, Lugano, Switzerland, Sep. 2019, pp. 1–9.
[162] T. Kornuta, D. Rajan, C. Shivade, A. Asseman, and A. S. Ozcan, "Leveraging medical visual question answering with supporting facts," in *Proc. Work. Notes CLEF Conf. Labs Eval. Forum*, Lugano, Switzerland, Sep. 2019, pp. 1–6.
[163] I. Allaouzi, M. B. Ahmed, and B. Benamrou, "An encoder–decoder model for visual question answering in the medical domain," in *Proc. Work. Notes CLEF Conf. Labs Eval. Forum*, Lugano, Switzerland, Sep. 2019, pp. 1–9.
[164] A. Al-Sadi, B. Talafha, M. Al-Ayyoub, Y. Jararweh, and F. Costen, "Just at ImageCLEF 2019 visual question answering in the medical domain," in *Proc. Work. Notes CLEF Conf. Labs Eval. Forum*, Lugano, Switzerland, Sep. 2019, pp. 1–6.
[165] A. B. Abacha, M. Sarrouti, D. Demner-Fushman, S. A. Hasan, and H. Müller, "Overview of the VQA-Med task at Imageclef 2020: Visual question answering and generation in the medical domain," in *Proc. Work. Notes CLEF Conf. Labs Eval. Forum*, Bucharest, Romania, 2021, pp. 1–9.
[166] M. Sarrouti, "NLM at VQA-Med 2020: Visual question answering and generation in the medical domain," in *Proc. Work. Notes CLEF Conf. Labs Eval. Forum*, Thessaloniki, Greece, Sep. 2020, pp. 1–6.
[167] H. Umada and M. Aono, "Kdevqa at VQA-med 2020: Focusing on GLU-based classification," in *Proc. Work. Notes CLEF Conf. Labs Eval. Forum*, Thessaloniki, Greece, Sep. 2020, pp. 1–9.
[168] S. Liu, H. Ding, and X. Zhou, "Shengyan at VQA-med 2020: An encoder–decoder model for medical domain visual question answering task," in *Proc. Work. Notes CLEF Conf. Labs Eval. Forum*, Thessaloniki, Greece, Sep. 2020, pp. 1–9.
[169] B. Jung, L. Gu, and T. Harada, "Bumjun_Jung at VQA-med 2020: VQA model based on feature extraction and multi-modal feature fusion," in *Proc. Work. Notes CLEF Conf. Labs Eval. Forum*, Thessaloniki, Greece, Sep. 2020, pp. 1–9.
[170] G. Chen, H. Gong, and G. Li, "HCP-MIC at VQA-med 2020: Effective visual representation for medical visual question answering," in *Proc. Work. Notes CLEF Conf. Labs Eval. Forum*, Thessaloniki, Greece, Sep. 2020, pp. 1–9.
[171] A. Al-Sadi, Hana'Al-Theiabat, and M. Al-Ayyoub, "The inception team at VQA-med 2020: Pretrained VGG with data augmentation for medical VQA and VQG," in *Proc. Work. Notes CLEF Conf. Labs Eval. Forum*, Thessaloniki, Greece, Sep. 2020, pp. 1–9.
[172] Q. Xiao, X. Zhou, Y. Xiao, and K. Zhao, "Yunnan university at VQA-med 2021: Pretrained biobert for medical domain visual question answering," in *Proc. Work. Notes CLEF Conf. Labs Eval. Forum*, Bucharest, Romania, Sep. 2021, pp. 1–9.
[173] J. Li and S. Liu, "Lijie at ImageCLEFmed VQA-med 2021: Attention model based on efficient interaction between multimodality," in *Proc. Work. Notes CLEF Conf. Labs Eval. Forum*, Bucharest, Romania, Sep. 2021, pp. 1–10.
[174] Y. Li, Z. Yang, and T. Hao, "TAM at VQA-med 2021: A hybrid model with feature extraction and fusion for medical visual question answering," Work. Notes CLEF, Tech. Rep., Sep. 2021.
[175] N. M. S. Sitara and S. Kavitha, "SSN MLRG at VQA-med 2021: An approach for VQA to solve abnormality related queries using improved datasets," in *Proc. CEUR Workshop*, Sep. 2021, 1329–1335.
[176] I. Chebbi, "VGG16: Visual generation of relevant natural language questions from radiology images for anomaly detection," Preprint Res. Gate, Tech. Rep., Sep. 2021.
[177] D. Gupta, S. Suman, and A. Ekbal, "Hierarchical deep multi-modal network for medical visual question answering," *Exp. Syst. Appl.*, vol. 164, Feb. 2021, Art. no. 113993.
[178] W. Zheng, L. Yan, F.-Y. Wang, and C. Gou, "Learning from the guidance: Knowledge embedded meta-learning for medical visual question answering," in *Proc. Int. Conf. Neural Inf. Process.*, Bangkok, Thailand, Cham, Switzerland: Springer, Nov. 2020, pp. 194–202.
[179] B. Liu, L.-M. Zhan, and X.-M. Wu, "Contrastive pre-training and representation distillation for medical visual question answering based on radiology images," in *Proc. 24th Int. Conf. Med. Image Comput. Comput. Assist. Intervent.–MICCAI*, Strasbourg, France, Cham, Switzerland: Springer, Sep. 2021, pp. 210–220.
[180] Y. Khare, V. Bagal, M. Mathew, A. Devi, U. D. Priyakumar, and C. Jawahar, "MMBERT: Multimodal BERT pretraining for improved medical VQA," in *Proc. IEEE 18th Int. Symp. Biomed. Imag. (ISBI)*, Nice, France, Apr. 2021, pp. 1033–1036.
[181] C. Dhanush, D. P. Kumar, and A. Kanavalli, "A VQA system for medical image classification using transfer learning," in *Data Engineering and Intelligent Computing: Proceedings of ICICC 2020*, Bengaluru, India, Cham, Switzerland: Springer, Sep. 2021, pp. 249–257.
[182] F. Cong, S. Xu, L. Guo, and Y. Tian, "Caption-aware medical VQA via semantic focusing and progressive cross-modality comprehension," in *Proc. 30th ACM Int. Conf. Multimedia*, New York, NY, USA, Oct. 2022, pp. 3569–3577.
[183] M. Wang, X. He, L. Liu, L. Qing, H. Chen, Y. Liu, and C. Ren, "Medical visual question answering based on question-type reasoning and semantic space constraint," *Artif. Intell. Med.*, vol. 131, Sep. 2022, Art. no. 102346.
[184] H. Wang, H. Pan, K. Zhang, S. He, and C. Chen, "M2FNet: Multi-granularity feature fusion network for medical visual question answering," in *Proc. PRICAI Trends Artif. Intell. 19th Pacific Rim Int. Conf. Artif. Intell.*, Shanghai, China, Cham, Switzerland: Springer, Non. 2022, pp. 141–154.
[185] T. Kornuta, "PyTorchPipe: A framework for rapid prototyping of pipelines combining language and vision," 2019, *arXiv:1910.08654*.
[186] A. Lubna, S. Kalady, and A. Lijiya, "MoBVQA: A modality based medical image visual question answering system," in *Proc. IEEE Region 10 Conf. (TENCON)*. Kochi, India: Institute of Electrical and Electronics Engineers, Oct. 2019, pp. 727–732.

[187] D. Sharma, S. Purushotham, and C. K. Reddy, "MedFuseNet: An attention-based multimodal deep learning model for visual question answering in the medical domain," *Sci. Rep.*, vol. 11, no. 1, pp. 1–18, Oct. 2021.

[188] A. Al-Sadi, M. Al-Ayyoub, Y. Jararweh, and F. Costen, "Visual question answering in the medical domain based on deep learning approaches: A comprehensive study," *Pattern Recognit. Lett.*, vol. 150, pp. 57–75, Oct. 2021.

[189] K. Gasmi, I. B. Ltaifa, G. Lejeune, H. Alshammari, L. B. Ammar, and M. A. Mahmood, "Optimal deep neural network-based model for answering visual medical question," *Cybern. Syst.*, vol. 53, no. 5, pp. 403–424, Jul. 2022.

[190] S. S. N. Mohamed and K. Srinivasan, "ImageCLEF 2020: An approach for visual question answering using VGG-LSTM for different datasets," in *Proc. Work. Notes CLEF Conf. Labs Eval. Forum*, Thessaloniki, Greece, Sep. 2020, pp. 1–9.

[191] H. T. Haridas, M. M. Fouda, Z. M. Fadlullah, M. Mahmoud, B. M. ElHalawany, and M. Guizani, "MED-GPVS: A deep learning-based joint biomedical image classification and visual question answering system for precision e-health," in *Proc. IEEE Int. Conf. Commun.*, Seoul, South Korea, May 2022, pp. 3838–3843.

[192] S. Al-Hadhrami, M. E. B. Menai, S. Al-Ahmadi, and A. Alnafessah, "An effective med-VQA method using a transformer with weights fusion of multiple fine-tuned models," *Appl. Sci.*, vol. 13, no. 17, p. 9735, Aug. 2023.

[193] Figshare. (2017). *Brain Tumor Dataset*. Accessed: Nov. 2023. [Online]. Available: https://figshare.com/articles/dataset/brain_tumor_dataset/1512427/5

[194] L. Seenivasan, M. Islam, G. Kannan, and H. Ren, "SurgicalGPT: End-to-end language-vision GPT for visual question answering in surgery," 2023, *arXiv:2304.09974*.

[195] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, "Visual7W: Grounded question answering in images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 4995–5004.

[196] D. Yu, J. Fu, T. Mei, and Y. Rui, "Multi-level attention networks for visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 4187–4195.

[197] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 1839–1848.

[198] I. Schwartz, A. Schwing, and T. Hazan, "High-order attention models for visual question answering," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–6.

[199] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 21–29.

[200] R. Li and J. Jia, "Visual question answering with question representation update (QRU)," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.

[201] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus, "Simple baseline for visual question answering," 2015, *arXiv:1512.02167*.

[202] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A neural-based approach to answering questions about images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, CL, USA, Dec. 2015, pp. 1–9.

[203] H. Xu and K. Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, Cham, Switzerland: Springer, Oct. 2016, pp. 451–466.

[204] M. Ren, R. Kiros, and R. Zemel, "Exploring models and data for image question answering," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.

[205] L. Ma, Z. Lu, and H. Li, "Learning to answer questions from image using convolutional neural network," in *Proc. 13th AAAI Conf. Artif. Intell.*, Arizona, USA, Feb. 2016, pp. 1–6.

[206] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu, "Are you talking to a machine? Dataset and methods for multilingual image question," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.

[207] I. Ilievski, S. Yan, and J. Feng, "A focused dynamic attention model for visual question answering," 2016, *arXiv:1604.01485*.

[208] P. Wang, Q. Wu, C. Shen, A. van den Hengel, and A. Dick, "Explicit knowledge-based reasoning for visual question answering," 2015, *arXiv:1511.02570*.

[209] P. Wang, Q. Wu, C. Shen, A. Dick, and A. van den Hengel, "FVQA: Fact-based visual question answering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 10, pp. 2413–2427, Oct. 2018.

[210] C. Xiong, S. Merity, and R. Socher, "Dynamic memory networks for visual and textual question answering," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 2397–2406.

[211] L. Yu, E. Park, A. C. Berg, and T. L. Berg, "Visual madlibs: Fill in the blank image generation and question answering," 2015, *arXiv:1506.00278*.

[212] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.

[213] C. Zhu, Y. Zhao, S. Huang, K. Tu, and Y. Ma, "Structured attentions for visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Honolulu, HI, USA, Oct. 2017, pp. 1300–1309.

[214] Q. Wu, C. Shen, P. Wang, A. Dick, and A. V. D. Hengel, "Image captioning and visual question answering based on attributes and external knowledge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1367–1381, Jun. 2018.

[215] A. Jabri, A. Joulin, and L. van Der Maaten, "Revisiting visual question answering baselines," in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9912. Berlin, Germany: Springer Verlag, 2016, pp. 727–739.

[216] V. Kazemi and A. Elqursh, "Show, ask, attend, and answer: A strong baseline for visual question answering," 2017, *arXiv:1704.03162*.

[217] X. Lin and D. Parikh, "Leveraging visual question answering for image-caption ranking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, Cham, Switzerland: Springer, Oct. 2016, pp. 261–277.

[218] H. Xu and K. Saenko, "Dual attention network for visual question answering," in *Proc. ECCV 2nd Workshop Storytelling Images Videos (VisStory)*, Amsterdam, The Netherlands, Oct. 2016, pp. 1–6.

[219] L. Cao, L. Gao, J. Song, X. Xu, and H. T. Shen, "Jointly learning attentions with semantic cross-modal correlation for visual question answering," in *Proc. Australas. Database Conf.*, Brisbane, QLD, Australia, Cham, Switzerland: Springer, Sep. 2017, pp. 248–260.

[220] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jun. 2018, pp. 6077–6086.

[221] J. Song, P. Zeng, L. Gao, and H. T. Shen, "From pixels to objects: Cubic visual attention for visual question answering," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Stockholm, Sweden, Jul. 2018, pp. 906–912.

[222] A. Osman and W. Samek, "Dual recurrent attention units for visual question answering," 2018, *arXiv:1802.00209*.

[223] C. Ma, C. Shen, A. Dick, Q. Wu, P. Wang, A. V. D. Hengel, and I. Reid, "Visual question answering with memory-augmented networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 6975–6984.

[224] V. Lioutas, N. Passalis, and A. Tefas, "Explicit ensemble attention learning for improving visual question answering," *Pattern Recognit. Lett.*, vol. 111, pp. 51–57, Aug. 2018.

[225] P. Gao, H. Li, S. Li, P. Lu, Y. Li, S. C. H. Hoi, and X. Wang, "Question-guided hybrid convolution for visual question answering," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 469–485.

[226] J. Liang, L. Jiang, L. Cao, L.-J. Li, and A. Hauptmann, "Focal visual-text attention for visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 6135–6143.

[227] D. Teney, P. Anderson, X. He, and A. V. D. Hengel, "Tips and tricks for visual question answering: Learnings from the 2017 challenge," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 4223–4232.

[228] L. Jiang, J. Liang, L. Cao, Y. Kalantidis, S. Farfade, and A. Hauptmann, "MemexQA: Visual memex question answering," 2017, *arXiv:1708.01336*.

[229] Z. Su, C. Zhu, Y. Dong, D. Cai, Y. Chen, and J. Li, "Learning visual knowledge memory networks for visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7736–7745.

[230] W. Zhang, J. Yu, H. Hu, H. Hu, and Z. Qin, "Multimodal feature fusion by relational reasoning and attention for visual question answering," *Inf. Fusion*, vol. 55, pp. 116–126, Mar. 2020.

[231] B. Sun, Z. Yao, Y. Zhang, and L. Yu, "Local relation network with multilevel attention for visual question answering," *J. Vis. Commun. Image Represent.*, vol. 73, Nov. 2020, Art. no. 102762.
[232] X. Zhu, Z. Mao, Z. Chen, Y. Li, Z. Wang, and B. Wang, "Object-difference drived graph convolutional networks for visual question answering," *Multimedia Tools Appl.*, vol. 80, no. 11, pp. 16247–16265, May 2021.
[233] S. Hashemi Hosseinabad, M. Safayani, and A. Mirzaei, "Multiple answers to a question: A new approach for visual question answering," *Vis. Comput.*, vol. 37, no. 1, pp. 119–131, Jan. 2021.
[234] Z. Bai, Y. Li, M. Woźniak, M. Zhou, and D. Li, "DecomVQANet: Decomposing visual question answering deep network via tensor decomposition and regression," *Pattern Recognit.*, vol. 110, Feb. 2021, Art. no. 107538.
[235] W. Zhang, J. Yu, Y. Wang, and W. Wang, "Multimodal deep fusion for image question answering," *Knowl.-Based Syst.*, vol. 212, Jan. 2021, Art. no. 106639.
[236] Y. Liu, X. Zhang, F. Huang, Z. Zhou, Z. Zhao, and Z. Li, "Visual question answering via combining inferential attention and semantic space mapping," *Knowl.-Based Syst.*, vol. 207, Nov. 2020, Art. no. 106339.
[237] Y. Liu, X. Zhang, Z. Zhao, B. Zhang, L. Cheng, and Z. Li, "ALSA: Adversarial learning of supervised attentions for visual question answering," *IEEE Trans. Cybern.*, vol. 52, no. 6, pp. 4520–4533, Jun. 2022.
[238] L. Gao, L. Cao, X. Xu, J. Shao, and J. Song, "Question-led object attention for visual question answering," *Neurocomputing*, vol. 391, pp. 227–233, May 2020.
[239] H. Zhong, J. Chen, C. Shen, H. Zhang, J. Huang, and X.-S. Hua, "Self-adaptive neural module transformer for visual question answering," *IEEE Trans. Multimedia*, vol. 23, pp. 1264–1273, 2021.
[240] J. Hong, S. Park, and H. Byun, "Selective residual learning for visual question answering," *Neurocomputing*, vol. 402, pp. 366–374, Aug. 2020.
[241] S. Lobry, D. Marcos, J. Murray, and D. Tuia, "RSVQA: Visual question answering for remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8555–8566, Dec. 2020.
[242] P. Lu, H. Li, W. Zhang, J. Wang, and X. Wang, "Co-attending free-form regions and detections with multi-modal multiplicative feature embedding for visual question answering," in *Proc. AAAI Conf. Artif. Intell.*, New Orleans, LA, USA, Feb. 2018, pp. 1–6.
[243] K. R. Chandu, M. A. Pyreddy, M. Felix, and N. N. Joshi, "Textually enriched neural module networks for visual question answering," 2018, *arXiv:1809.08697*.
[244] J. Singh, V. Ying, and A. Nutkiewicz, "Attention on attention: Architectures for visual question answering (VQA)," 2018, *arXiv:1803.07724*.
[245] L. Peng, Y. Yang, Y. Bin, N. Xie, F. Shen, Y. Ji, and X. Xu, "Word-to-region attention network for visual question answering," *Multimedia Tools Appl.*, vol. 78, no. 3, pp. 3843–3858, Feb. 2019.
[246] A. S. Toor, H. Wechsler, and M. Nappi, "Question action relevance and editing for visual question answering," *Multimedia Tools Appl.*, vol. 78, no. 3, pp. 2921–2935, Feb. 2019.
[247] D. Teney and A. van den Hengel, "Visual question answering as a meta learning task," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 219–235.
[248] D. Yu, X. Gao, and H. Xiong, "Structured semantic representation for visual question answering," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Athens, Greece, Oct. 2018, pp. 2286–2290.
[249] M. Malinowski, C. Doersch, A. Santoro, and P. Battaglia, "Learning visual question answering by bootstrapping hard attention," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 3–20.
[250] N. Ruwa, Q. Mao, L. Wang, J. Gou, and M. Dong, "Mood-aware visual question answering," *Neurocomputing*, vol. 330, pp. 305–316, Feb. 2019.
[251] Y. Niu, K. Tang, H. Zhang, Z. Lu, X.-S. Hua, and J.-R. Wen, "Counterfactual VQA: A cause-effect look at language bias," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12695–12705.
[252] N. Ouyang, Q. Huang, P. Li, Y. Cai, B. Liu, H.-F. Leung, and Q. Li, "Suppressing biased samples for robust VQA," *IEEE Trans. Multimedia*, vol. 24, pp. 3405–3415, 2022.
[253] C. Yang, S. Feng, D. Li, H. Shen, G. Wang, and B. Jiang, "Learning content and context with language bias for visual question answering," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2021, pp. 1–6.
[254] Z. Liang, H. Hu, and J. Zhu, "LPF: A language-prior feedback objective function for de-biased visual question answering," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 1955–1959.
[255] R. Cadene, C. Dancette, M. Cord, and D. Parikh, "RUBi: Reducing unimodal biases for visual question answering," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.
[256] D. Yuan, "Language bias in visual question answering: A survey and taxonomy," 2021, *arXiv:2111.08531*.

SUHEER AL-HADHRAMI received the master's degree in computer science from King Saud University, in 2016, where she is currently pursuing the Ph.D. degree with the Department of Computer Science. She is an Assistant Teacher with Hadhramout University. Her research interests include computer vision, NLP, machine learning, and assistive technology.

MOHAMED EL BACHIR MENAI received the Ph.D. degree in computer science from the Mentouri University of Constantine, Algeria, and the University of Paris VIII, France, in 2005, and the Ph.D. degree Habilitation Universitaire in computer science from the Mentouri University of Constantine, in 2007 (it is the highest academic qualification in Algeria, France, and Germany). He is currently a Professor with the Department of Computer Science, King Saud University. His main research interests include satisfiability and optimization problems, natural language processing, machine learning, and AI in medicine.

SAAD AL-AHMADI is currently an Associate Professor with the Department of Computer Science, College of Computer and Information Sciences, King Saud University, Saudi Arabia. Also, he is a part-time consultant in many agencies. He has published many papers in reputable journals and conferences. His current research interests include AI for healthcare, the IoT security, and adversarial machine learning.

AHMED ALNAFESSAH is the general manager of the Smart Cities Technologies Institute at King Abdulaziz City for Science and Technology (KACST). He is also an AI Lead in the Centre for the C4IR KSA in Affiliation with the World Economic Forum WEF. He was a Senior AI and Cloud Computing Engineer at the National Centre for AI and Big Data Technologies, KACST, from 2009 to 2017. His research interests include performance engineering for big data systems and AI, specifically on in-memory platforms. He is interested in cognitive city systems, digital twins, big data, AI, the IoT, HPC, and complex distributed systems. He has a strong background and experience in applying AI solutions for architecture design, continuous integration, and continuous delivery (ci/cd), DevOps testing tools, and runtime services management with international (Europe, U.K., and Silicon Vally) and local experience. He was a team member who developed the AI DevOps framework called RADON. This DevOps framework helps the European software industry to adopt serverless function as a service (FaaS) technology while avoiding lock-in within a specific FaaS provider by utilizing AI/ML and DevOps.