A Survey on Visual Question Answering Methodologies

Aya M. Al-Zoghby[1, 2], Aya Salah Saleh[3] and Wael Abdelkader Awad[4]
[*1]Computer Science Department, Faculty of Computers and Artificial Intelligence, Damietta University, New Damietta, 34517, Egypt
[1]aya_el_zoghby@du.edu.eg
[3]aya.saleh92@gmail.com
[4] wael_abdelkader@du.edu.eg
[**2]Computer Science and Engineering faculty, New Mansoura University, New Mansoura, 35511, Egypt
[2]aya.elzohby@nmu.edu.eg

Abstract: ***Understanding visual question-answering (VQA) will be essential for many human tasks. However, it poses significant obstacles at the core of artificial intelligence as a multimodal system. This article provides a summary of the challenges in multimodal architectures that have lately been demonstrated by the enormous rise in research. We need to keep our eyes on these challenges to enhance the design of visual question-answering systems. Then we will introduce the recent rapid developments in methods for answering visual questions with images. Providing the right response to a natural language question concerning an input image, it is a difficult multi-modal activity as we don't need only to extract features from both modal (text and image) but also getting attention on relation between them. Many deep learning researchers are drawn to it because of their outstanding contributions to text, voice, and vision technologies (images and videos) in fields like welfare, robotics, security, and medicine, etc.***

Key words: ***Deep Learning, Visual question answering, Multimodal challenges, VQA methodologies.***

1 INTRODUCTION

Artificial intelligence (AI) is essential for emulating humans' ability to see, touch, smell, hear, and feel [1].
We can merge these senses as multimodal by process and link this information together using deep learning to build a multimodal system (MMS). Multimodal deep learning (MMDL) is used to merge different types of modalities to enhance the learning experience. As unimodal learning cannot cover all human learning aspects, MMDL allows various senses to be merged in information analysis and processing to achieve better understanding. We can combine multiple types of modalities (video, audio, body gestures, physiological signals, facial expressions, text, and images) with multiple intelligence algorithms to introduce different higher performance styles of learning because we differ in the ways we learn. The input can be a multiple modality and the output will be another form of modal, or we can use one modality to assist another [2].
To present MMDL, we will combine heterogeneous data from multiple resources to allow more accurate prediction [3]. The heterogeneity of data and interconnection of modalities brings a unique challenge to understanding these challenges, as shown in Fig. 1 [2, 4]:

1- Representation: several pieces of information will be presented from different media that contain redundant and complementary data. We must deal with missing data, noise, and learn representations that reflect cross-modal interactions between individual elements and cross different modalities. This is the core building block for most multimodal model problems, which have sub-challenges such as:
- Fusion: the idea of integrating information from different modalities.
- Coordination: where the number of representations of the modalities is equal and each modal retains each space.
- Fission: not only integrating information but also integrating multiple facts that reflect knowledge about internal structure (factorization, variant, clustering).

2- Alignment: identifying the cross-modal connection between all elements of multiple modalities by measuring the similarities among different media and aligning the long-term dependence as most modalities have an internal structure with multiple elements per sample; therefore, some sub-challenges will appear, such as:
- Discrete alignment: connections between discrete elements.
- Continuous alignment: connections between ambiguous segmentation and continuous signals.
- Contextualized representation: captures cross- modal interaction between elements.

Therefore, we can deduce that the goal of alignment is that each word aligns with each object with respect to the representation to maintain the structure.

3- Reasoning: combining knowledge through multiple inferential steps and exploiting the multimodal alignment problem structure. Reasoning sub problems involves:
- Structure model: how to model this structure.
- Intermediate concept: like conventional representation, which is a composition process, parameterization is studied.
- Inference paradigm: abstract concepts from individual multimodal evidence.

- External knowledge: This is the core challenge in MM, which studies the definition of composition and structure.

4- Generation: used to produce raw modalities that reflect cross-modal interaction structure and coherence. Its sub-challenges are as follows:
- Summarization: highlight important information and reduce information from one modality to another that is similar.
- Translation: maintaining consistent information from one modality to another.
- Creation: understanding the modalities and expanding the knowledge.

5- Transference: used to help noisily target modalities by transferring knowledge between them. We can categorize its sub-challenges into:
- Cross-modal transfer: where one modal adapts and modifies the representation of the other.
- Co-learning: transfer information and knowledge from a rich modality to help a noisy modal.
- Modal induction: motivates behavior in both modalities and keeps them separate.

6- Qualification: empirical and theoretical studies to better understand heterogeneity cross-modal interaction and the multimodal learning process, which involves the following:
- Heterogeneity: data are very heterogeneous.
- Interaction: how can we visualize and understand the type of interaction?
- Learning: understanding the learning that occurred in these models.

VQA is a MML model that merges natural language processing (NLP), computer vision (CV), and relational reasoning by asking questions related to visual information in video, image, and audio to make reasoning and produce an answer for natural language questions. To solve this problem, numerous VQA and visual reasoning techniques have been developed. On the other hand, every known method relies on extracting features from both the question and the image before combining them to produce an answer, so it considers a classification problem to predict the answer. Using deep learning is essential to solve more complex machine learning problems such as VQA problems by training models using large amounts of data and complex algorithms. Several techniques are used in VQA, such as (LSTM- Bi LSTM- transformers - CNN – fast CNN – graph representation) to extract text and image features for the VQA model. To obtain an accurate answer to a specific question according to our image Not only extracting features is required but we also need to compare the semantics of information present in both modalities and obtain the relationship between objects and important regions in the image and question together. The way these methods combine textual and visual data is where they diverge most from one another. Concatenate them, for instance, and subject them to a linear classifier. Alternative methods include the use of Bayesian models to show the fundamental connections between the distributions of the question, image, and response features. This section examines several contemporary architectural proposals for VQA and visual reasoning tasks. We divide these models into the following basic categories: traditional neural network techniques, external knowledge, transformers, attention mechanism, graph, and neural-symbolic. We go into further depth about VQA methodologies in the sections that follow.

2 LITERATURE REVIEW

A. VQA using Traditional Neural Network Models

In early methods, CNN to extract image features and LSTM to extract question features were used, sending them in a single embedding, then combining these feature vectors (image- text) together to obtain image– question feature representation in common space. This technique, which makes representation in common feature space called joint embedding, uses image captioning [5-7] and image notation [8].

Another model called “VIS-LSTM” which used LSTM and CNN to embed visual semantic features as input, was proposed [9]. Then “iBOWIMG” model was suggested as a straightforward VQA model that uses a bag of words as a text feature. The input question was converted to word feature by word embedding and image feature extraction by CNN, which is a multiclass logistic regression model (classification model) that predicts the answer [10]. An mQA model was produced using LSTM to extract question representation and store the linguistic context in the answer, and CNN was used to extract visual representation [11]. A model focused on open-ended and multiple-choice questions was produced, and image features were extracted using faster region-based CNN (R-CNN) and LSTM for question feature extraction. In this model, R-CNN is 13% more accurate than any other image recognition model by updating the weight matrix of R-CNN according to the question weight [12]. “N-KBSN” model based on joint embedding using dynamic word vectors was improved using a faster R-CNN for image feature extraction and an ELMo model for text characterization and feature enhancement based on the multi-head attention mechanism, which are the three basic parts of N-KBSN [13].

B. VQA using External Knowledge based Models

As existing databases do not contain all real-world occurrences and activities, external knowledge is crucial in real-world circumstances. VQA tasks perform better when knowledge base (KB) databases are linked. Several external KB approaches for VQA tasks have been proposed throughout the DL era. Marino, K. et al. [14] provided a novel dataset for VQA in which the answers to the questions depend on outside information sources. More than 14,000 questions are

included. The categories in this dataset include sports, history, science, and technology. Instead of just comprehending the question and image attributes, this dataset necessitates the use of other resources. Yu, J. et al. [15] introduced a graph-based recurrent reasoning network (GRUC) for visual question answering that requires outside information. This network focuses on cross-modal knowledge reasoning on graph-structured multimodal knowledge representations. Various knowledge graphs collect image information from three perspectives (semantic, factual, and visual), and recurrent reasoning models extract the necessary representation from various model spaces. The Knowledge Reasoning with Implicit and Symbolic Representations (KRISP) approach was divided into two categories: knowledge representations and reasoning. In the beginning, implicit knowledge could be efficiently learned from unsupervised language pertaining and supervised training data using transformer-based models. Second, knowledge bases are explicitly and symbolically encoded. This strategy combined these two by integrating symbolic representations from a knowledge network while never losing their explicit semantics to an implicit embedding and taking advantage of the potent implicit reasoning of transformer models for response prediction [16]. Song, D., et al. [17] developed a novel Knowledge Enhanced Visual-and-Linguistic BERT (KVL-BERT) model to include commonsense knowledge into the visual-and-linguistic BERT, which can enhance cognition-level visual understanding and reasoning skills. In addition to collecting input from visual and linguistic content, the multi-layer Transformer incorporates external commonsense information taken from ConceptNet. They suggested an algorithm termed RMGSR (Relative-position-embedding and Mask-self-attention Guided Semantic Representations) to maintain the structural details and semantic representation of the original text. A multimodal semantic graph knowledge reasoning model (MSG-KRM) was presented to perform reasoning and deep fusion of image– text data with outside knowledge sources. The process of creating the semantic graph involves extracting keywords from the question text, external knowledge texts, and image object detection information, which are then represented as symbol nodes. The knowledge graph was then used to build three different types of semantic graphs: vision, query, and the external knowledge text. Non-symbol nodes were then added to connect these three distinct graphs, and they were each annotated with the appropriate node and edge types. In the inference phase, a type-aware graph attention module is used for deep reasoning after the multimodal semantic graph and image– text data are embedded into the feature semantic network using three different embedding techniques [18]. The Dual Scene Graph Enhancement Module (DSGEM), which introduces explicit relational reasoning during intra-modal feature interaction, was proposed. Visual and textual scene graphs were created for this purpose using common sense and syntactic structure. In addition, the textual scene graph provided grammatical relationships between words, enriching the semantic relationships between visual objects with commonsense relationships. Then, two scene graph improvement submodules are suggested to actively direct the feature interaction between objects (nodes) by propagating the pertinent scene graph structure information. The explicit reasoning capacity was added using DSGEM to the existing VQA models [19].

C. VQA using Attention Mechanism Models

It is noticed that there are two drawbacks on traditional approach:

- First, they ignore the relationship between objects or regions to get accurate prediction and focus only on particular regions in images or salient semantic objects.
- Second, predicting the current answer according to the most relevant object or image region in the current time step ignoring the previous object or image region [20].

Therefore, numerous VQA models use an attention mechanism that focuses on visually important content in the image. In VQA and visual reasoning, various attention models have been developed to allow learning attention over visual, textual, or both modalities [21-24].

The attention model is used to overcome noisy and unnecessary information during the prediction phase. When extracting features from different regions, it specifies the priorities of these features. VQA models focused on specific regions in the image according to the asked question. Peng, W., et al. [25] presented a co-attention method to generate reasoning parts using a joint image and question facts according to higher order. In a related study, bottom-up top-down attention that calculates the attention of object and image regions, which increases VQA model accuracy, was proposed [26]. YU, Z., et al. [27] proposed a deep modular co-attention network (MCAN) for a VQA task in which a modular co-attention (MCA) layer consists of self-attention questions and image representation. Guo, W., et al. [28] developed a re-attention model for VQA. They calculated the similarities between objects in the feature space to combine the image and the question. Then, the associated object of the image is re-attender to reconstruct the attention map to generate an answer. A VQA model based on attention from images and its relationship with semantic description of questions to focus on target areas using word vectors was developed [29]. A selective residual learning (SelRes) framework, which is a self-attention based VQA module that depends on the attention map and limits residual learning, indicating the importance of the input vector, was developed to improve performance. They also proposed adaptive SelRes that removes the heuristic selection rate by determining the selection through the network [30]. Guo, W., et al. [31] presented a re-attention model for VQA using observation from questions and answers to describe visual contents to guide visual attention learning and attention loss to evaluate the difference between visual attention learned by only questions and that learned in re-attention. They calculated the similarities for each object and re-attended the visual objects of the image according to the answer. Li, Q., et al. [32] proposed a VQA model with external image features and used an internal attention mechanism. This model differs from other models because the image and question are input, but the image is external. In this model, the features of the image were extracted using a question-oriented attention mechanism to

perform feature fusion. This model is based on adversarial learning and bidirectional attention. A Scene Graph-based co-Attention Network (SceneGATE) for Text VQA was proposed to clarify the relationships among objects, optical character recognition (OCR) tokens, and question words. To capture the intra-modal interaction between language and visuals as a guide for inter-modal interactions, we developed a guided attention module. Two attention modules, a scene graph-based semantic relation-aware attention and a positional relation-aware attention, were proposed and combined to make teaching the relationships between the two modalities obvious [33].

D. VQA using Graph and Neural-Symbolic models.

A graph neural network is used to capture the visual relationship between two or more regions that are important in an image or semantic object in an input image. Li, L., et al. [34] proposed a model for VQA called the relation- aware graph attention network (ReGAT). In this model, the visual features of the image are encoded into a graph, and for learning question adaptive representation, they use multi-type inter-object relations via a graph attention mechanism. Both explicit and implicit relations for visual object relations were used to represent geometric position and semantic interaction between objects and capture hidden dynamic between image regions. A bilinear graph network (BGN) was developed to model the context of the joint embedding of words and objects. In this model, there were two types of graph corpses: image- graph enabling the output nodes to have semantic and information by transferring features of the detected objects to their query words and question graph, which exchange information from image graph output nodes to get the relation between objects. They used the VQA v2.0 dataset, which proved the efficiency of the model in handling complex questions [35]. Sharma, H. and Jalal, A. [36] implemented a graph neural network (GNN) model using the relationship between significant regions or semantic objects in an image. The VQA model also employs an attention model that remembers previously attended information with the current object or region in the image using contextual LSTM. They used two datasets, VQA 1.0 and VQA 2.0. In which results illustrate that the proposed VQA model predicts more accurate answers and performs better. Xiong, P., et al. [37] proposed a multimodal technique (SA-VQA) that captures the deep connection between visual and textual modalities using graph representation of visual and textual content by converting visual and textual entities into sequential graphs and then merging them in structured alignment to improve reasoning performance. They used the GQA dataset and the non-per-trained VQA-v2 dataset.

E. VQA using Transformer Models

Theoretically, the Transformer has a strong capacity for reasoning. Transformers have been the de facto architecture in natural language processing (NLP) and have driven the development of numerous sequence prediction tasks. The prevailing methods first model long-range dependencies between words in generic corpora before making sense of sequence modeling and transduction tasks, such as language modeling and machine translation. Many researchers have attempted to further advance the transformer wave in computer vision (CV) to capitalize on the enormous success of NLP. [38].

Models based on transformers perform exceptionally well in visual question answering (VQA). However, their performance suffers when we measure them on the basis of systematic generalization, which involves handling novel combinations of well-known concepts. Neural module networks (NMNs), which are composed of modules or neural networks that focus on a particular subtask, are a potential method for systematic generalization. Inspired by NMNs and Transformers [39]. Self-Adaptive Neural Module Transformer (SANMT), which considers intermediate Q&A outcomes to adapt to changes in both question feature encoding and layout decoding. To be more precise, they used a unique transformer module to encode the intermediate outcomes with the provided question features to create a dynamic question feature embedding those changes as the reasoning progressed. In addition, the transformer uses the intermediate outcomes of each reasoning stage to direct the layout arrangement of the following step [40]. Koshti, D., et al. [41] used bottom-up features and a different co-attention mechanism to create a unique BERT-based hierarchical VQA model. Using BERT and a hierarchical approach for the joint representation of image and question features, the proposed model enhances the question feature extraction module. They proved that bottom-up features with the BERT-based hierarchical approach outperform both the simple hierarchical VQA model and the top-down bottom approach for VQA. A new transformer network named the dual-decoder transformer network (DTNN) was proposed, which predicts linguistic answers and visual instances simultaneously using two decoupled decoders. In this model, a set of instance query embedding and a sampling strategy were introduced around the mass center to precisely locate the object. They divided the image into bins so that complex coordinate regression could be transformed into a simpler classification prediction. The model parameters and computational costs can be significantly decreased by integrating image region characteristics with grid data, enabling DDTN to be trained and inferred on low-cost devices [42]. The GFTransformer for aerial image VQA was suggested as a novel model with gated attention modules and a mutual fusion module, which could provide a starting point for further study. To create a thorough baseline for computer vision and earth observation research, they suggested a new VQA dataset for high-resolution aerial photos, called HRVQA. With 10 different types of questions, they introduced a semi-automatic construction scheme for labeling the image/question pairs [43].

F. Other Models

It is noticed that there are several approaches for VQA, so here we will present some different studies used for VQA. YU, Z., et al. [44] performed a multimodal factorized bilinear (MFB) pooling approach using a co-attention mechanism to

jointly learn images and questions. The bilinear pooling-based model has a higher dimensional representation with high computational complexity, but it is better than traditional linear modes. In another study, ViQAR (Visual Question Answering and Reasoning) was introduced, where a multi-word answer is generated for a visual query. The model must generate the complete answer and a rationale that seeks to justify the generated answer. An end-to-end architecture to solve this task and describe how to evaluate it was proposed. This model produced strong answers and rationales through qualitative and quantitative evaluation and the human Turing Test examination [45].

The most well-known vision-language tasks, Visual Question Answering (VQA) and Image Captioning (CAP), have equivalent scene-text variants that call for inference from the text in the image in question. The two are addressed separately despite their evident similarities, leading to task-specific approaches that can either be seen or read, not both. Therefore, captioning models have been designed for large language models. For example, researchers designed PROMPTCAP (Prompt-guided image Captioning), which was trained using examples with GPT-3, a question-aware captioning model that uses natural language to describe visual content. This model helps GPT-3 achieve better results in the OK-VQA and AOKVQA datasets. [46]. Ganz, R., et al. [47] carefully examined that topic and provided UniTNT, a unified text-non-text method, which gives existing multimodal systems scene-text understanding capabilities. To be more precise, we consider scene-text data as an additional modality and integrate it into any pretrained encoder decoder-based architecture via defined modules. Extensive testing demonstrated that UniTNT produces the first single model that can manage both types of tasks.

3 CONCLUSION

Visual question answering is considered an AI-complete task because of the power of deep learning methods that combine computer vision and natural language processing to create visual dialog. It is a difficult problem that needs resolving subtasks like activity recognition, object detection, spatial relationships between objects and common-sense reasoning. In this paper, we investigate and present a comprehensive survey on QA, showing the challenges in MMDL (Fig. 1) and focusing on QA methodologies, as shown in Table 1. VQA is a challenging study area that requires algorithms with the ability to perform challenging recognition and reasoning tasks. We analyze each methodology and determine the data set used. In future work, researchers should deal with synthetic data and sophisticated reasoning to deal with world problems that require external knowledge.

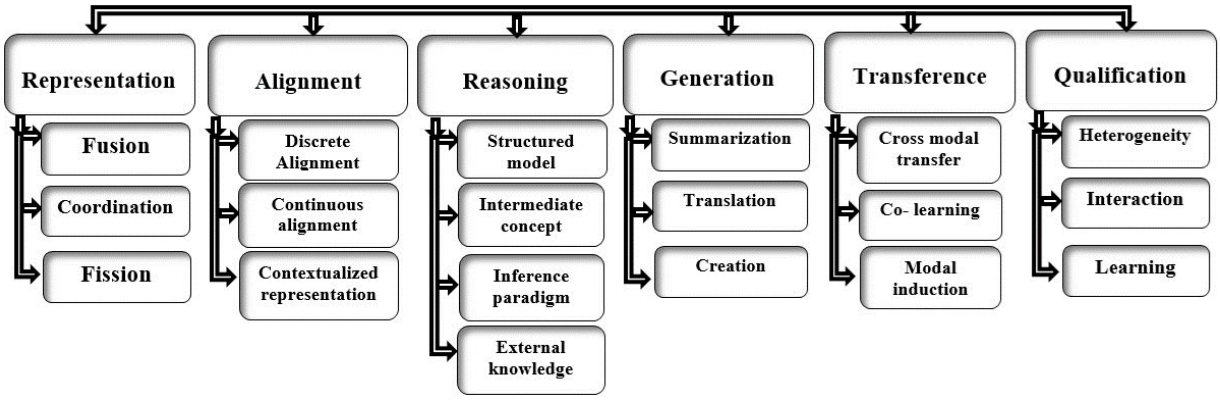

Fig. 1 Challenges in MMDL

TABLE I
METHODS AND MODELS SUMMARY

Methodology	Model	year	Question feature	Image feature	dataset
Traditional Neural Network Models	VIS-LSTM [9]	2015	CNN	LSTM	DAQUAR and COCO-QA
	iBOWIMG [10]	2015	CNN	Word embedding	COCO VQA
	mQA [11]	2015	CNN	LSTM	FM-IQA
	Open ended- MCQ [12]	2019	Faster R-CNN	LSTM	DAQUAR and COCO-QA
	N-KBSN [13]	2021	Faster R-CNN	Elmo	VQA2.0, MS-COCO
External Knowledge based	OK-VQA [14]	2019	RNN/ GRU	CNN/ResNet	OK-VQA
	GRUC [15]	2020	RNN/LSTM	Faster-RCNN,	FVQA, OKVQA, Visual7W+KB
	KRISP [16]	2021	BERT	Faster R-CNN	OK-VQA
	KVL-BERT [17]	2021	BERT	Faster R-CNN	VCR
	MSG-KRM [18]	2023	BERT	Faster R-CNN	OK-VQA
	DSGEM [19]	2023	BERT / GloVe	Faster R-CNN	VQA V2, OK-VQA
Attention Mechanism	Peng.W et al [25]	2017	RNN/LSTM	CNN/ VGG or ResNet	Visual Genome and VQA real
	Bottom up down attention [26]	2018	RNN/ GRU	Faster R-CNN.	Visual Genome
	(MCAN) [27]	2019	RNN/LSTM	Faster R-CNN/ResNet	VQA-v2
	Guo.W et al [28]	2020	RNN/LSTM	Faster R-CNN	VQA v2
	Xi.Y et al. [29]	2020	LSTM	CNN	DQAUAR, COCO-QA
	(SelRes) [30]	2020	LSTM/ GloVe	Faster R-CNN/ VGG or ResNet	VQA 2.0, MS-COCO
	GUo.W et al. [31]	2020	LSTM	Faster R-CNN	VQA v2
	Li.Q et al. [32]	2021	GRU/ Glove	Faster-RCNN	VQA2.0
	SceneGATE [33]	2023	Bert	Faster-RCNN	Text-VQA and ST-VQA
Graph and Neural-Symbolic	(ReGAT) [34]	2019	RNN/LSTM	Faster R-CNN	VQA 2.0 and VQA-CP v2
	BGN [35]	2023	LSTM/ Transformer (BERT)	Faster-RCNN	VQA v2.0
	(GNN) [36]	2021	contextual LSTM (GRU)	ResNet	VQA 1.0 and VQA 2.0.
	(SA-VQA) [37]	2022	Glove/ MLP	Faster R-CNN	GQA, VQA-v2
Transformer	Zhong.H, et al [40]	2021	bidirectional LSTM/ transformer	ResNet- two CNN layers/ transformer	CLEVR, CLEVR-CoGenT, VQAv1.0, and VQAv2.0
	Koshti.D et al [41]	2022	BERT	Faster R-CNN/ ResNet	VQA v2.0
	DTNN [42]	2023	LSTM/ transformer encoder in dual encoder	object detector	VizWiz Ground and GQA
	Li, K. et al [43]	2023	Glove-LSTM/ Transformer encoder	ResNet/ Transformer decoder	RSVQA-HR/ HRVQA
Another models	(MFB) [44]	2017	RNN/ LSTM	CNN/ ResNet	MS-COCO
	ViQAR [45]	2021	Bert/ LSTM	spatial attention model	VCR
	Ganz, R., et al [47]	2023	OCR Encoder	VL-OCR Decoder	VQAv2, Text VQA, ST-VQA, COCO Captions and Text Caps

REFRENCES

[1] T. Baltrušaitis, C. Ahuja and L. -P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423-443, 2019.

[2] S. Jabeen, X. Li, M. S. Amin, O. Bourahla, S. Li and A. Jabbar "A review on methods and applications in multimodal deep learning," *ACM Transactions on Multimedia Computing, Communications, and Applications,* Vol.19, no. 2, pp. 1–41, 2023.

[3] S. Sengupta, S. Basak, P. Saikia, S. Paul, V. Tsalavoutis, F. Atiah, V. Ravi and A. Peters, "A Review of Deep Learning with Special Emphasis on Architectures, Applications and Recent Trends," *Knowledge Based System*, vol. 194, p. 105596, 2020.

[4] P.P. Liang, A. Zadeh and L-P. Morency, (2023) "Foundations and Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions," Available from: https://arxiv.org/abs/2209.03430v2, (accessed 20 Feb 2023).

[5] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko and T. Darrell, "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," *IEEE Trans Pattern analysis and machine intelligence*, vol. 39, no. 4, pp. 677-691, 2017.

[6] J. Mao, W. Xu, Y. Yang, J. Wang and A.L. Yuille, "Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN)," *in Proc. of 3rd International Conference on Learning Representations, {ICLR}*, San Diego, CA, USA, May 2015.

[7] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, "Show and tell: A neural image caption generator," *in Proc. of 2015 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 3156-3164, Boston, MA, USA, June 2015.

[8] W. Zhang, H. Hu and H. Hu, "Training Visual-Semantic Embedding Network for Boosting Automatic Image Annotation," *Neural Processing Letters*, vol. 48, no.3, pp.1503–1519, 2018.

[9] M. Ren, R. Kiros and R. S. Zemel, "Exploring Models and Data for Image Question Answering," *in Proc. of the 28th International Conference on Neural Information Processing Systems (NIPS'15)*, vol. 2, pp. 2953–2961, Montreal, Quebec, Canada, December 2015.

[10] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam and R. Fergus, (2015) "Simple Baseline for Visual Question Answering," Available from: https://arxiv.org/abs/1512.02167v2 , (accessed 15 Dec 2015).

[11] 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. of the 28th International Conference on Neural Information Processing Systems (NIPS'15)*, vol.2, pp. 2296-2304, Montreal, Quebec, Canada, December 2015.

[12] S. Jha, A. Dey, R. Kumar and V. K.-Solanki, "A Novel Approach on Visual Question Answering by Parameter Prediction using Faster Region Based Convolutional Neural Network," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol.5, no. 5, pp. 30-37, 2019.

[13] Z. Ma, W. Zheng, X. Chen and L. Yin, "Joint embedding VQA model based on dynamic word vector," *PeerJ. Computer Science*, vol. 7, no. 353, pp. 1–20, 2021.

[14] K. Marino, M. Rastegari, A. Farhadi and R. Mottaghi, "OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge," *in Proc. of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3190-3199, Long Beach, California, USA, June 2019.

[15] J. Yu, Z. Zhu, Y. Wang, W. Zhang, Y. Hu and J. Tan, "Cross-modal knowledge reasoning for knowledge based visual question answering," *Pattern Recognition*, vol. 108, no. 10, p. 107563, 2020.

[16] K. Marino, X. Chen, D. Parikh, A. Gupta and M. Rohrbach," KRISP: Integrating Implicit and Symbolic Knowledge for Open-Domain Knowledge-Based VQA," *in Proc. of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14106-14116, Nashville, TN, USA, June 2021.

[17] D. Song, S. Ma, Z. Sun, S. Yang and L. Liao, "KVL-BERT: Knowledge Enhanced Visual-and-Linguistic BERT for visual commonsense reasoning," *Knowledge-Based Systems*, vol. 230, p. 107408, 2021.

[18] L. Jiang and Z. Meng, "Knowledge-Based Visual Question Answering Using Multi-Modal Semantic Graph," *Electronics*, vol. 12, no. 6, pp. 1-19, 2023.

[19] B. Wang, Y. Ma, X. Li, H. Liu, Y. Hu and B. Yin, "DSGEM: Dual scene graph enhancement module-based visual question answering," *IET Computer Vision*, vol. 17, no. 6, pp. 638-651, 2023.

[20] W. Zhang, J. Yu, W. Zhao and C. Ran, "DMRFNET: Deep multimodal reasoning and fusion for visual question answering and explanation generation," *Information Fusion*, vol. 72, no. 3, pp. 70-79, 2021.

[21] Z. Yang, X. He, J. Gao, L. Deng and A. Smola, "Stacked Attention Networks for Image Question Answering," *in Proc. of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 21-29, Las Vegas, NV, USA, June 2016.

[22] J. Lu, J. Yang, D. Batra and D. Parikh, "Hierarchical Question-Image Co-Attention for Visual Question Answering," *in Proc. of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*, vol. 30, pp. 289-297, Barcelona, Spain, December 2016.

[23] V. Kazemi and A. Elqursh (2017), "Show, Ask, Attend, and Answer: A Strong Baseline for Visual Question Answering," Available from: https://arxiv.org/abs/1704.03162v1, (accessed 11 Apr 2017).
[24] D. Nguyen and T. Okatani, "Improved Fusion of Visual and Language Representations by Dense Symmetric Co-attention for Visual Question Answering," *in Proc. of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6087–6096, Salt Lake City, UT, USA, June 2018.
[25] P. Wang, Q. Wu, C. Shen and A. Hengel, "The vqa-machine: Learning how to use existing vision algorithms to answer new questions," *in Proc. of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.3909-3918, Honolulu, Hawaii, USA, July 2017.
[26] 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. of the 2018 IEEE conference on computer vision and pattern recognition*, pp. 6077–6086, Salt Lake City, UT, USA, June 2018.
[27] Z. Yu, J. Yu, Y. Cui, D. Tao and Q. Tian, " Deep modular co-attention networks for visual question answering," *in Proc. of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 6281–6290, Long Beach, California, USA, June 2019.
[28] W. Guo, Y. Zhang, X. Wu, J. Yang, X. Cai and X. Yuan, " Re-Attention for Visual Question Answering," *in Proc. of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 91-98, New York, USA, February 2020.
[29] Y. Xi, Y. Zhang, S. Ding and S. Wan, "Visual question answering model based on visual relationship detection," *Signal Processing: Image Communication,* vol. 80, p. 115648, 2020.
[30] J. Hong, S. Park and H. Byun, "Selective residual learning for visual question answering," *Neurocomputing*, vol. 402, no. 10, pp. 366-374, 2020.
[31] W. Guo, Y. Zhang, X. Wu, J. Yang, X. Cai and X. Yuan, " Re-Attention for Visual Question Answering," *in Proc. of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 91-98, New York, USA, February 2020.
[32] Q. Li, X. Tang and Y. Jian, " Adversarial Learning with Bidirectional Attention for Visual Question Answering," *Sensors*, vol. 21, no. 21, pp. 7164, 2021.
[33] F. Cao, S. Luo, F. Núñez, Z. Wen, J. Poon and S. C. Han, "SceneGATE: Scene-Graph based co-Attention networks for Text visual question answering," *Robotics*, vol. 12, no.4, pp. 114, 2023.
[34] L. Li, Z. Gan, Y. Cheng and J. Liu, "Relation-aware graph attention network for visual question answering," *in Proc. of IEEE/CVF International Conference on Computer Vision (ICCV*), pp. 10312–10321, Seoul, Korea (South), November 2019
[35] D. Guo, C. Xu and D. Tao. "Bilinear Graph Networks for Visual Question Answering," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 2, pp.1023-1034, 2023.
[36] Sharma and Jalal A., "Visual question answering model based on graph neural network and contextual attention", *Image and Vision Computing*, vol. 110, no. 1, pp. 104-165, 2021.
[37] P. Xiong, Q. You, P. Yu, Z. Liu and Y. Wu, (2022), "SA-VQA: Structured Alignment of Visual and Semantic Representations for Visual Question Answering," Available from: https://arxiv.org/abs/2201.10654, (accessed 25 Jan 2022).
[38] Y. Cheng, Z. Zhao, Z. Wang and H. Duan, "Rethinking vision transformer through human–object interaction detection," *Engineering Applications of Artificial Intelligence*, vol. 122, p.106123, 2023.
[39] M. Yamada, V. D'Amario, K. Takemoto, X. Boix and T. Sasaki, (2023), "Transformer Module Networks for Systematic Generalization in Visual Question Answering," Available from: https://arxiv.org/abs/2201.11316v2 , (accessed 17 Mar 2023).
[40] Zhong, J. Chen, C. Shen, H. Zhang, J. Huang and X. S. Hua, "Self-Adaptive Neural Module Transformer for Visual Question Answering," *IEEE Transactions on Multimedia*, vol. 23, pp. 1264-1273, 2021.
[41] D. Koshti, A. Gupta, and M. Kalla, "BERT based Hierarchical Alternating Co-Attention Visual Question Answering using Bottom-Up Features," *Int J Intell Syst Appl Eng*, vol. 10, no. 3, pp. 158–168, 2022.
[42] L. Zhu, L. Peng, W. Zhou and J. Yang, "Dual-decoder transformer network for answer grounding in visual question answering," *Pattern Recognition Letters*, vol. 171, no. 7, pp. 53-60, 2023.
[43] Li, G. Vosselman and M. Y. Yang, (2023) "HRVQA: A Visual Question Answering Benchmark for High-Resolution Aerial Images," https://arxiv.org/abs/2301.09460v1, (accessed 23 Jan 2023).
[44] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," *in proc. Of the IEEE international conference on computer vision (ICCV),* pp. 1839-1848, Venice, Italy, October 2017.
[45] R. Dua, S.S. Kancheti, and V.N. Balasubramanian, "Beyond VQA: Generating Multi-word Answers and Rationales to Visual Questions," *in proc. Of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1623-1632, Nashville, TN, USA, June 2021.
[46] Y. Hu, H. Hua, Z. Yang, W. Shi, N.A. Smith, and J. Luo, (2023) "PromptCap: Prompt-Guided Task-Aware Image Captioning," Available from: https://arxiv.org/abs/2211.09699v4, (accessed 17 Aug 2023).
[47] R. Ganz, O. Nuriel, A. Aberdam, Y. Kittenplon, S. Mazor, and R. Litman, (2023) "Towards Models that Can See and Read," Available from: https://arxiv.org/abs/2301.07389v2 , (accessed 21 Mar 2023).

BIOGRAPHY

Aya M. El-Zoghby
Associate Professor at Computer Science Department in faculty of computer and artificial intelligence at Damietta University and an Associate Professor at Computer Science and Engineering faculty, New Mansoura University.

Aya Salah Saleh
Received the B.Sc. degree from Mansoura University, Egypt, in 2013, and a researcher at M.Sc. degree at Damietta University.

Wael Abdelkader Awad
Professor at Computer Science Department in faculty of computer and artificial intelligence at Damietta University.

ARABIC ABSTRACT

استطلاع حول منهجيات الإجابة على الأسئلة المرئية

آية محمد الزغبي[1, 2]، آية صلاح صالح[3]، ووائل عبد القادر عوض[4]

[1] *قسم علوم الحاسب، كلية الحاسبات والذكاء الاصطناعي، جامعة دمياط، دمياط الجديدة، 34517، مصر*
aya_el_zoghby@du.edu.eg[1]
aya.saleh92@gmail.com[3]
wael_abdelkader@du.edu.eg [4]
[2] *كلية علوم وهندسة الحاسبات، جامعة المنصورة الجديدة، المنصورة الجديدة، 35511، مصر*
aya.elzohby@nmu.edu.eg[2]

الملخص: ***سيكون فهم الإجابة على الأسئلة المرئية ضروريًا للعديد من المهام البشرية. ومع ذلك، فإنه يشكل عقبات كبيرة في جوهر الذكاء الاصطناعي كنظام متعدد الوسائط. تقدم هذه المقالة ملخصًا للتحديات التي تواجه البنى متعددة الوسائط والتي تم إثباتها مؤخرًا من خلال الارتفاع الهائل في الأبحاث. نحن بحاجة إلى إبقاء أعيننا على هذه التحديات لتحسين تصميم أنظمة الإجابة على الأسئلة المرئية. ثم سنعرض التطورات السريعة الحديثة في طرق الإجابة على الأسئلة المرئية بالصور. إن تقديم الإجابة الصحيحة على سؤال باللغة الطبيعية يتعلق بصورة كمُدخل، يعد نشاطًا صعبًا متعدد الوسائط، حيث لا نحتاج فقط إلى استخراج الميزات من كل من الوسائط (النص والصورة) ولكن نحتاج أيضًا إلى جذب الانتباه إلى العلاقة بينهما. ينجذب العديد من الباحثين في مجال التعلم العميق للإجابة على الأسئلة المرئية بسبب مساهماتهم البارزة في تقنيات النصوص والصوت والرؤية (الصور ومقاطع الفيديو) في مجالات مثل الرعاية الاجتماعية والروبوتات والأمن والطب وما إلى ذلك.***

الكلمات المفتاحية: ***التعلم العميق، الإجابة على الأسئلة المرئية، تحديات الوسائط المتعددة، منهجيات الأسئلة المرئية.***