ArabicQuest: Enhancing Arabic Visual Question Answering with LLM Fine-Tuning

Asmaa ElMaghraby, Samaa Maged, Mohamed Essawey, Rawan ElFaramawy, Esraa Negm, Ghada Khoriba
Artificial Intelligence Program,
School of Information Technology and Computer Science (ITCS),
Nile University, Giza, Egypt
{a.elmaghraby, sa.maged, m.abdelmaged, r.elfaramawy, e.negm, ghadakhoriba}@nu.edu.eg

***Abstract*—In an attempt to bridge the semantic gap between language understanding and visuals, Visual Question Answering (VQA) offers a challenging intersection of computer vision and natural language processing. Large Language Models (LLMs) have shown remarkable ability in natural language understanding; however, their use in VQA, particularly for Arabic, is still largely unexplored. This study aims to bridge this gap by examining how well LLMs can improve VQA models. We use state-of-the-art AI algorithms on datasets from multiple fields, including electric devices, Visual Genome, RSVQA, and ChartsQA. We introduce ArabicQuest, a Text Question Answering (TQA) tool that combines Arabic inquiries with visual data. We assess the performance of LLMs across various question types and image settings and find that fine-tuning methods such as LLaMA-2, BLIP-2, and Idefics-9B-Instruct models provide encouraging results, although challenges still arise in counting and comparison tasks. Our findings demonstrate the importance of advancing VQA further—especially for Arabic—to enhance accessibility and user satisfaction in a variety of applications.**

***Index Terms*—Visual Question Answering(VQA),Text Question Answering(TQA),Large Language Models(LLMs), ArabicQuest**

I. Introduction

In the dynamic realm of human-computer interaction, a fascinating fusion of text-based question-answering and vision-driven comprehension has given birth to Visual Question Answering (VQA) systems. These systems face the monumental task of enabling machines to seamlessly understand and respond to questions about visual content, effectively bridging the gap between images and natural language. This interdisciplinary pursuit delves into analyzing textual queries and interpreting visual stimuli, presenting various technical challenges and innovative solutions. Visual Question Answering (VQA) has sparked interest in computer vision and natural language processing research. The VQA task involves answering a question about an image, using textual embedding from questions, and bridging the semantic disparity between image and question [1].

Over the years, researchers have delved into numerous Visual Question Answering (VQA) approaches, from conventional feature-based methods to contemporary deep-learning techniques. However, the emergence of Large Language Models (LLMs) marks a significant turning point in natural language processing (NLP). Leveraging self-attention mechanisms and extensive pre-training on extensive text datasets, LLMs have broken through linguistic barriers, demonstrating remarkable capabilities in processing natural language text across various contexts.

In this research, our approach lies in developing ArabicQuest: Visual Question to Actions modeling for Enhanced Arabic Interface Systems. This ambitious initiative seeks to address the unique challenges posed by VQA in Arabic, endeavoring to blend the subtle nuances of Arabic text with the complexities of visual comprehension. Through rigorous experimentation and evaluation, we aim to uncover the synergies between text and vision in the context of VQA while advancing accessibility and user satisfaction across diverse linguistic landscapes. The distinct challenges presented by the Arabic language in VQA, especially when leveraging Large Language Models (LLMs), further underscore the significance of our research endeavor.

Through rigorous experimentation and evaluation, we aim to uncover the synergies between text and vision in the context of VQA while advancing accessibility and user satisfaction across diverse linguistic landscapes. The distinct challenges presented by the Arabic language in VQA, especially when leveraging Large Language Models (LLMs), further underscore the significance of our research endeavor.

The rest of this paper is organized as follows: Section 2 presents Related Work, Section 3 shows the Methodology, Section 4 Results, and Section 5 concludes.

II. Related Work

VQA has emerged as a popular and captivating area of research, garnering significant interest from researchers in recent years. Creating VQA datasets and their approaches presents a challenging task, as these datasets typically consist of images and associated questions and answers. In this section, we offer an overview of the approaches of VQA datasets that have been most frequently utilized in the reviewed papers.

A. English Visual Question Answering

Innovative approaches have been proposed in recent developments in English Visual Question Answering. Vanilla VQA [2], regarded as a standard for evaluating deep learning techniques on VQA dataset [3], utilizes CNN for feature extraction and LSTM or recurrent networks for language interpretation. These extracted features are merged using element-wise operations into a unified feature and then used to classify

one of the answers. Teney et al. [4] pioneered the integration of object detection into VQA models and emerged victorious in the VQA dataset [3]. The model aids in narrowing down features and applying enhanced attention to images. Leveraging the R-CNN architecture, this model demonstrated remarkable accuracy improvements over other architectures.

Stacked Attention Networks [5] introduced attention by utilizing the softmax output of the intermediate question feature on the DAQUAR dataset [6]. The attention among the features is stacked, enabling the model to focus on crucial parts of the image. Neural-Symbolic VQA [7] designed for the CLEVR datasetCLEVR [8], this model utilizes the question formation and image generation strategies inherent to CLEVR. It transforms images into structured features and converts question features into their original root question strategy. These features are then employed to filter out the required answer.

Focal Visual Text Attention [9] applied attention based on both text components and ultimately classifies the features needed to respond to the question. This model is particularly suitable for VQA in videos, encompassing more diverse use cases than images. Differential Networks [10] on the COCOQA [11] leverage the distinctions between forward propagation steps to mitigate noise and learn the interdependencies among features. Image features are extracted using Faster-RCNN [12]

Kim et al. [13]introduced a semantic scene graph generation method grounded in the Resource Description Framework (RDF) paradigm on the FVQA dataset [14]. This approach aims to elucidate semantic connections within scenes. To construct the scene graph, they utilize Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), which are deep learning models. Wu et al. [15] proposed a method for image-related questions by integrating image content representations with knowledge base information. They used a CNN to generate attribute-based representations from the image, creating internal image captions. SPARQL accessed relevant knowledge from an external database based on predicted attributes, with Doc2vec capturing semantic meaning from retrieved knowledge paragraphs. These attributes, captions, and knowledge vectors were fed into an LSTM to predict the answer to input questions in a word sequence.

Zheng et al. [16] proposed a novel approach called cross-modal mixture experts (CMMEs) as the fusion encoder for the RSIVQA dataset. CMMEs consist of visual and textual experts, replacing the conventional feed-forward networks in standard transformers. Each CMME block contains experts specialized in capturing fusion information from their respective modalities, allowing for the dynamic switching between modal experts. Lobry et al. [17] introduced the task of Visual Question Answering (VQA) from remote sensing images as a versatile method for extracting information from remotely sensed data. Their method involves building datasets for VQA that can be extended and adapted to different data sources. They proposed two datasets, RSVQA-HR and RSVQA-LR, representing different resolutions, and they serve as valuable resources for further research in the field of VQA from remote sensing imagery.

Singh and Shekar [18] proposed the LEAF-QA dataset by incorporating public metadata from charts and introduced a transformers-based framework leveraging chart structural properties. Extensive experimentation showcased the framework's state-of-the-art performance on recent Chart Q/A datasets. They conducted experiments with pre-training tasks, demonstrating their benefits for the Chart Q/A problem. Through attention mechanisms, they dissected the model to understand answer retrieval functionality.

B. Arabic Visual Question Answering

Kamel et al. [19] created the first dataset and system for Visual Question Answering in Arabic (VAQA). The dataset is fully automatically generated and contains 5000 real-world images taken from the MS-COCO dataset, 2712 unique questions resulting in 137,888 Image Question Answers. The dataset is divided into 60%, 20%, and 20% for training, validation, and testing, respectively. The proposed Arabic-VQA system consists of five modules: question pre-processing, textual features extraction, visual features extraction, feature fusion from both modalities and answer prediction.

Arabic VQA still lacks extensive research and development compared to its English counterpart. Our study aims to bridge this gap by exploring further advancements in Arabic VAQA, opening up new avenues for understanding and interacting with visual content in the Arabic language.

III. Methodology

This study focuses into two fields of artificial intelligence: Text Question Answering (TQA) and Visual Question Answering (VQA), focusing on processing both natural languages and visual data. Our approach is the development of ArabicQuest, a system engineered to offer question-answering-driven task mostly aimed to Arabic-speaking Middle Eastern users, this is done by using Blip [20], Blip2 [21], LLaMa [22], and Idefics-9B-Instruct model [23].

A. System Description

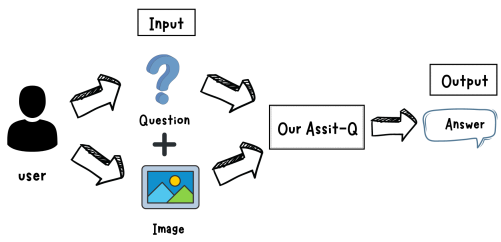

Fig. 1. Context diagram of the interaction with the system.

ArabicQuest, as shown in Figure 3, is designed to seamlessly integrate visual data with natural language inquiries, leveraging advanced AI and NLP techniques to provide users

with detailed instructions for interacting with electronic devices and answering questions related to satellite images, charts, and other visual content. ArabicQuest underscores the importance of practical applications in the Arabic-speaking community, striving for ease of use and affordability. by combining the exploration of Text Question Answering (TQA) and Visual Question Answering (VQA) in this unique context.

B. Text Question Answering (TQA)

1) Dataset Description: This dataset was generated from an extensive review of approximately 20 microwave (manually collected) machine catalogs using AI tools. It consists of 378 Arabic questions and answers, with categories including operation, allowed foods, and Safety Precautions.

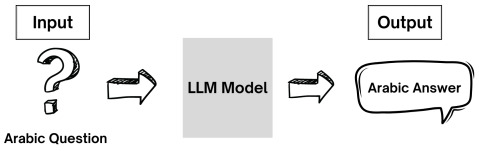

Fig. 2. The TQA Model Processing Pipeline.

2) Model Description and Implementation: The LLaMA Large Language Model [22] was chosen for its advanced natural language processing abilities. This model has two versions 7B and 65B parameters and is adept at interpreting Arabic queries and generating relevant answers, as illustrated in Figure 2.

Regarding hyper-papers of the model, it was fine-tuning focusing on adapting to the ArabicQuest system. Involving key parameter adjustments: LoRA set to 64 for enhanced self-attention, LoRA Alpha at 16 for learning rate control, and a LoRA Dropout rate of 0.1 to prevent overfitting.

C. Visual Question Answering (VQA)

1) Dataset Description:

a) Visual Genome Dataset: The Visual Genome dataset [24] is a rich collection that helps the exploration of connection of using language and vision. With 108,077 images and 1.7 million questions and answers, it goes beyond simple object recognition. This dataset makes machines look at pictures and answer questions about everything from ”what” is in an image to ”where” things are, ”when” something is happening, ”who” is in the picture, ”why” something is there, and ”how” things are occurring. These questions require machines to have a deeper understanding of the context, not just recognize things. Additionally, a part of this dataset gives detailed descriptions of objects, their features, and how they relate to each other in some images. This makes the Visual Genome dataset a great tool for researchers who want to improve and bridge the gap of machine understanding and interacting with visual information and language [25].

b) Remote Sensing Visual Question Answering Dataset (RSVQA): 11,431 questions and answers are included with 772 low-resolution photos from the Sentinel-2 satellite in the RSVQA Dataset [17]. This dataset is used to solve issues with object identification and recognition, using images which are inherently complicated. Through the complexities of remote sensing data images, the RSVQA aims to evaluate and enhance AI's capacity to comprehend and interpret comprehensive satellite images in order to extract valuable information. [26].

c) ChartsQA Dataset: A total of 23.1k rows in the ChartQA Dataset, used for question answering and understanding all the details of a chart images [27]. By providing a large collection of chart images and questions that AI models may be trained to answer, the ChartQA Dataset aims to advance the area of visual data usage. Developing of systems that can comprehend and extract data from a variety of charts shapes and styles is made easier by it, which significantly advances the field of visual question-answering research. [28].

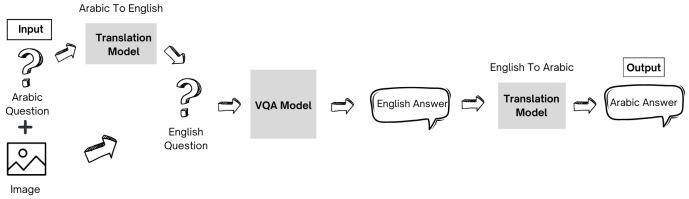

Fig. 3. The VQA Model Processing Pipeline.

2) Model Description and Implementation: Our VQA pipeline presents an innovative approach to using images to answer questions. The process starts with an Arabic question, subsequently translated into English using the Google Translate model, as seen in Figure 3. The input for the LLM is the English question and the associated image. After producing an English response, the system translates it back into Arabic so the user can receive the final response in Arabic. This pipeline demonstrates our dedication to improving accessibility and overcoming language barriers by guaranteeing that consumers receive accurate information in Arabic.

a) Visual Genome Dataset: We used the BLIP [20] (Bootstrapped Language Image Pretraining) model for the Visual Genome dataset, which was optimized over three epochs on a subset of the dataset (1000 rows). With this method, we could take full advantage of BLIP's envisioning and question-understanding skills and adapt them to our dataset's complex relationships between text and visual data.

We also applied the Idefics-9B-Instruct model [23], which was only optimized on 50% of the dataset for a single epoch, in addition to the BLIP model. This solution aims to enhance the system's capacity to handle complicated queries linked to the dataset's rich visual and textual content by utilizing Ideafics-9B-Instruct's skills in processing complex language instructions.

b) RSVQA Dataset: The Remote Sensing Visual Question Answering used the BLIP2 [21] model. This model modification was trained on our dataset across six epochs to optimize object recognition and identification in the difficult field of remote sensing.

c) ChartsQA Dataset: We used BLIP2 LLM [21] for the ChartsQA Dataset, fine-tuning it for a single epoch on a chosen subset of the dataset (1500 rows). This approach aimed to enhance the model's capability to comprehend chart images and gain insights based on data from visual representations.

IV. Results

Our experiments employed a large dataset encompassing nearly 2.7 million image-question-answer triplets sourced from various datasets. Table I presents a detailed analysis of our findings across different datasets and models. For remote sensing datasets, the models achieved an accuracy of approximately 64%. However, limitations were observed in handling counting tasks, suggesting that the model might require improvements, especially for count-related tasks. ChartQA reached 40% accuracy. The IDEFIC model exhibited better performance than Blip-2 on the visual egocentric dataset, achieving an accuracy of 50% and 43%, respectively. These results indicate the potential need for more training epochs or a more complex model architecture to enhance Blip-2's effectiveness on this specific dataset. Additionally, the llama 2 model achieved 70% accuracy on text question answering. Figures 4,5, 6, 7depicting the performance of each model

TABLE I
Model's Results

Model	Dataset	No. Epochs	Validation Loss	Accuracy
Llama	Electric Devices	1	0.9236	69%
Blip-2	RSVQA-LR	6	1.1722	64%
Blip-2	ChartsQA	1	2.1451	40%
Blip	Visual Genome	3	2.0397	43%
IDEFICS-9b	Visual Genome	3	1.0495	50%-

Fig. 4. IDEFIC model performance on Visual genome dataset

Fig. 5. Blip2 model performance on Visual Genome dataset

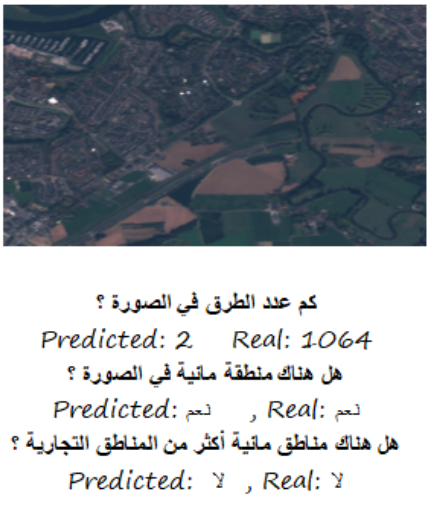

Fig. 6. Blip2 model performance on RSVQA-LR dataset

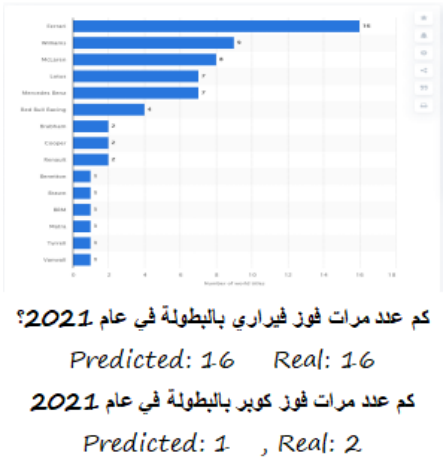

Fig. 7. Blip2 model performance on ChartsQAdataset

V. Conclusion

This study explored the field of Visual Question Answering (VQA), specifically aiming to enhance the performance of

Large Language Models (LLMs) when processing Arabic text in VQA scenarios. The research highlighted the critical role of LLMs such as BLIP and LLaMA in advancing VQA systems tailored for Arabic applications. By refining the ArabicQuest system, this work demonstrates that LLMs are adept at processing natural language queries and effectively integrating visual information. The findings illustrate the capacity of LLMs to connect the linguistic and visual realms, although challenges remain in specific areas like numerical counting and comparative questions. Overall, this research contributes to the advancement of VQA technologies and underscores the importance of their application in contexts relevant to the Arabic-speaking world.

REFERENCES

[1] A. A. Yusuf, C. Feng, X. Mao, R. Ally Duma, M. S. Abood, and A. H. A. Chukkol, "Graph neural networks for visual question answering: a systematic review," *Multimedia Tools and Applications*, pp. 1–38, 2023.

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

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

[4] D. Teney, P. Anderson, X. He, and A. Van Den Hengel, "Tips and tricks for visual question answering: Learnings from the 2017 challenge," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4223–4232, 2018.

[5] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 21–29, 2016.

[6] M. Malinowski and M. Fritz, "Towards a visual turing challenge," *arXiv preprint arXiv:1410.8027*, 2014.

[7] K. Yi, J. Wu, C. Gan, A. Torralba, P. Kohli, and J. Tenenbaum, "Neural-symbolic vqa: Disentangling reasoning from vision and language understanding," *Advances in neural information processing systems*, vol. 31, 2018.

[8] J. Johnson, B. Hariharan, L. Van Der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

[9] J. Liang, L. Jiang, L. Cao, L.-J. Li, and A. G. Hauptmann, "Focal visual-text attention for visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6135–6143, 2018.

[10] C. Wu, J. Liu, X. Wang, and R. Li, "Differential networks for visual question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 8997–9004, 2019.

[11] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755, Springer, 2014.

[12] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[13] S. Kim, T. H. Jeon, I. Rhiu, J. Ahn, and D.-H. Im, "Semantic scene graph generation using rdf model and deep learning," *Applied Sciences*, vol. 11, no. 2, p. 826, 2021.

[14] P. Wang, Q. Wu, C. Shen, A. Dick, and A. Van Den Hengel, "Fvqa: Fact-based visual question answering," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 10, pp. 2413–2427, 2017.

[15] Q. Wu, P. Wang, C. Shen, A. Dick, and A. Van Den Hengel, "Ask me anything: Free-form visual question answering based on knowledge from external sources," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4622–4630, 2016.

[16] X. Zheng, B. Wang, X. Du, and X. Lu, "Mutual Attention Inception Network for Remote Sensing Visual Question Answering," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

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

[18] H. Singh and S. Shekhar, "Stl-cqa: Structure-based transformers with localization and encoding for chart question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3275–3284, 2020.

[19] S. M. Kamel, S. I. Hassan, and L. Elrefaei, "Vaqa: Visual arabic question answering," *Arabian Journal for Science and engineering*, vol. 48, no. 8, pp. 10803–10823, 2023.

[20] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*, pp. 12888–12900, PMLR, 2022.

[21] D. Zhu, J. Chen, K. Haydarov, X. Shen, W. Zhang, and M. Elhoseiny, "Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions," *arXiv preprint arXiv:2303.06594*, 2023.

[22] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[23] K. Ahrabian, Z. Sourati, K. Sun, J. Zhang, Y. Jiang, F. Morstatter, and J. Pujara, "The curious case of nonverbal abstract reasoning with multi-modal large language models," *arXiv preprint arXiv:2401.12117*, 2024.

[24] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, pp. 32–73, 2017.

[25] D. A. Chacra and J. Zelek, "The topology and language of relationships in the visual genome dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4860–4868, 2022.

[26] S. Lobry, B. Demir, and D. Tuia, "Rsvqa meets bigearthnet: a new, large-scale, visual question answering dataset for remote sensing," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pp. 1218–1221, IEEE, 2021.

[27] P. Kavehzadeh, "Chart question answering with an universal vision-language pretraining approach," 2023.

[28] A. Masry, D. X. Long, J. Q. Tan, S. Joty, and E. Hoque, "Chartqa: A benchmark for question answering about charts with visual and logical reasoning," *arXiv preprint arXiv:2203.10244*, 2022.