Nile University جامعة النيل الأهلية | SCHOOL OF INFORMATION TECHNOLOGY & COMPUTER SCIENCE

ArabicQuest: Visual Question-to-Actions Modeling for Enhanced Arabic Interfaces

Asmaa El-Maghraby 019200036

Esraa Negm 202000799

Mohamed Essawey 202000440

Rawan El-Faramawy 202001762

Samaa El-Emary 202000597

Supervisor:

Assoc. Prof. Ghada Ahmed

February 2024

1. Refined Project Description.

A notable gap exists in defining well-structured tasks and establishing appropriate benchmarks in the domain of intelligent assistants, particularly within affordance-centric real-world scenarios. These intelligent assistants, including AR glasses, have long been designed to support everyday tasks, such as "How can I run the microwave for 1 minute?" There is a lack of clarity regarding task definitions and assessment criteria, specifically in Arabic. We aim to develop a new task and dataset called Visual Question-to-Actions for Enhanced Arabic interfaces, which aims to assist users in real-world scenarios through AI assistants and promote the development of more effective and user-friendly assistants for Arabic speakers. The task involves answering user questions about performing specific actions on everyday objects, such as appliances, using natural language queries and visual information.

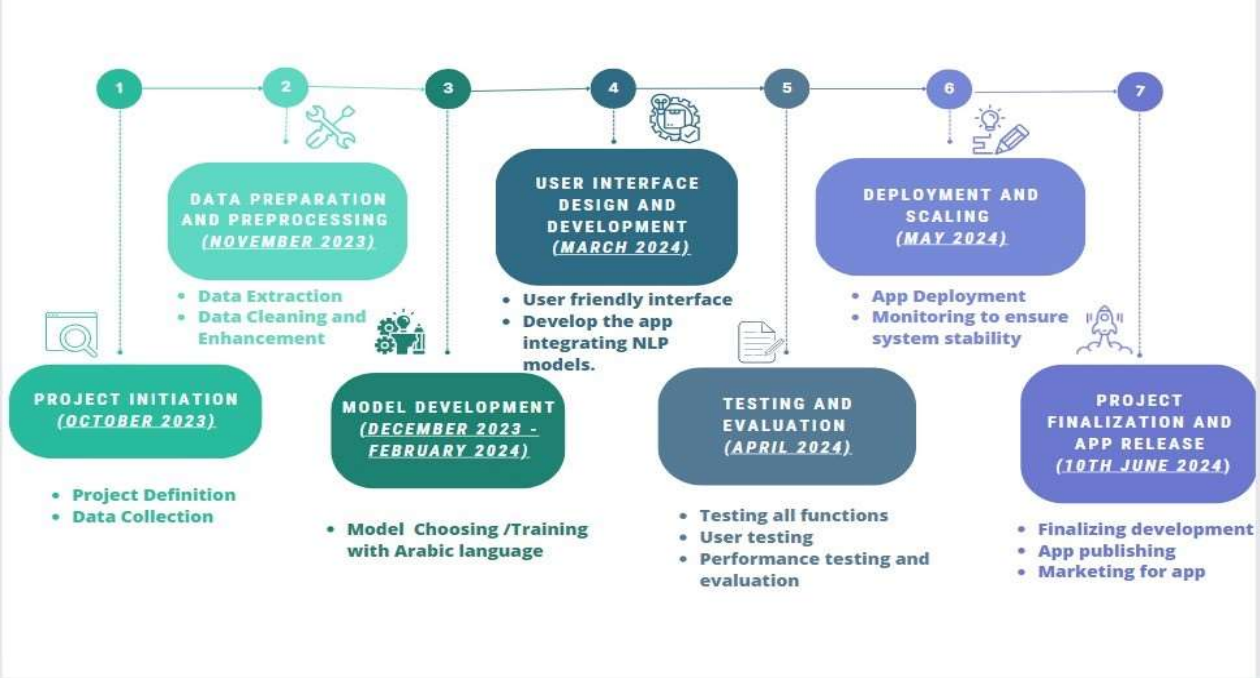
Our innovative Arabic-AssistQ project extends beyond merely assisting users with operating devices or appliances. By incorporating remote sensing data and the visual genome dataset, we are broadening the scope of our application to cater to diverse needs and interests of Arabic-speaking users. With this expansion, Arabic-AssistQ becomes a comprehensive tool for accessing information and guidance across various domains. Imagine being able to utilize Arabic-AssistQ not only to navigate unfamiliar devices but also to explore and understand the world around you. With remote sensing dataset (RSVQA), users can access valuable information about geographical features, environmental conditions, and spatial dynamics in Arabic. This feature empowers users to explore and learn about their surroundings with ease and convenience.

Moreover, the integration of the visual genome dataset adds another dimension to Arabic-AssistQ, allowing users to delve into the realm of visual understanding and interpretation. Users can now interact with images, objects, and scenes in Arabic, gaining insights and knowledge about visual content with the help of our application.

Furthermore, the incorporation of augmented reality (AR) technology elevates the user experience to a whole new level. Through real-time translations and language assistance, users can seamlessly interact with their environment, whether it's scanning real-world objects, deciphering text, or exploring images. This immersive and interactive approach makes learning and discovery more engaging and visually stimulating for Arabic-speaking users. In essence, Arabic-AssistQ transcends the traditional boundaries of language assistance and device operation. It empowers Arabic-speaking users to explore, learn, and interact with their surroundings and visual content in their native language, creating a truly enriching and empowering experience.

2. Refined Project Plan.

2.1 Milestone schedule:



2.2 Roles and Responsibilities distribution:

Name	Role	Role description
Dr. Ghada Khoriba	Project supervisor	supervisor is participating in the project planning process; coordinate labor needs and guides us. provide technical leadership for every aspect of the project. involved in strategic plans to accomplish business goals. collaborate effectively with all team members as well as hold regular team meetings.

Asmaa Mohamed	Team Leader	Documentation and LLMs/NLP model development.
Esraa Negm	Team Member	UI/UX (mobile application) and data collection.
Mohamed Abdelmaged	Team Member	Presentation and LLMs/NLP model development.
Samaa Maged	Team Member	Marketing/business model and LLMs/NLP model development
Rawan Mohamed	Team Member	Unit Testing / final testing and data preprocessing

2.3 Risk management and Contingency plans

2.3.1 Risk management:

- Data availability: There is a lack of high-quality Arabic language data that can be used to train NLP and LLMs models.
- Ambiguity and Polysemy: Arabic language often exhibits ambiguity and polysemy, where a word or phrase can have multiple meanings. The system needs to handle these

linguistic challenges effectively to provide accurate and contextually appropriate answers.

- Data Privacy and Security: Handling user queries and potentially sensitive information requires robust data privacy and security measures. Adequate safeguards should be implemented to protect user data from unauthorized access or misuse.

- Continuous Improvement and Updates: Arabic-AssistQ system should be regularly maintained, updated, and improved to stay relevant and accurate. Risks associated with system updates, bug fixes, and version control need to be managed effectively to prevent service disruptions or regression in performance.

- System Performance and Scalability: The system's performance, in terms of response time and scalability, is critical for a smooth user experience. Potential risks include system bottlenecks, high latency, or inability to handle a large volume of concurrent user queries.

2.3.2 Contingency plans:

- Data availability: Data augmentation for the lack of high-quality Arabic language data, we employ data augmentation techniques. This involves generating synthetic data by applying various transformations and modifications to existing data.

- Ambiguity and Polysemy: Develop contextual understanding mechanisms within the system to identify, resolve ambiguous queries by taking the surrounding context into account and provide multiple

possible answers with confidence scores to indicate uncertainty when dealing with ambiguous queries.

- Data Privacy and Security: Employ robust encryption techniques to protect user data during transmission and storage. Also, implement strict access controls and regularly conduct security audits to identify and address any vulnerabilities or risks to user data.

- Continuous Improvement and Updates: Establish a process for regular system updates and improvements, including version control and rollback plans in case of unexpected regressions or issues.

- System Performance and Scalability: Conduct load testing and performance tuning to ensure the system can handle high user volumes and queries without significant performance degradation.

3. System Requirements.

3.1 Requirements Elicitation Process

a) Description of the processes that were used.

The processes user for requirements elicitation mainly focused first on Surveys. To evaluate what people, think of a text-based question-answering system for household machine inquiries, we distributed an Arabic-language survey. User experiences with household appliances, and the frequency of problems experienced were all gathered through the survey. This allowed us to determine the level of demand for our system and identify the typical issues that users run into. To better understand and analyse the collected data, we

constructed a flow diagram. This diagram fig.2 visually represented the common user journey using our website.

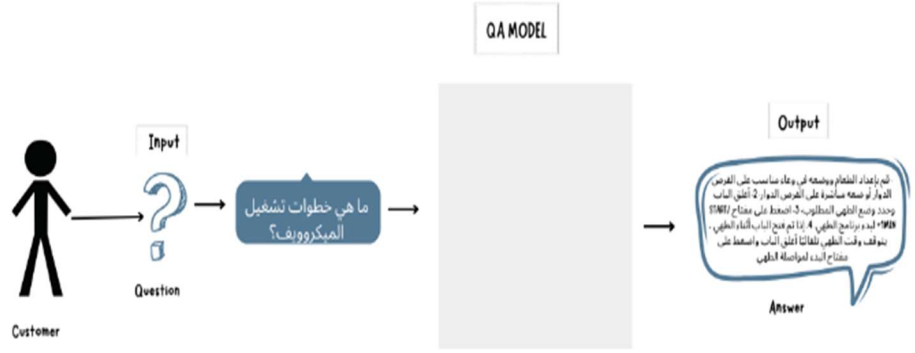


Fig.2 The flow diagram of the text QA.

Fig.3. flow diagram for the Visual Genome dataset.

Fig.4. flow diagram for the RSVQA dataset.

These flow diagrams depict the sequence of operations from the user's perspective, starting from input to the final output. They were crucial during brainstorming sessions to understand and map out the process flow, ensuring the system's design meets user requirements for both the Visual Genome Fig 3 and RSVQA datasets Fig 4.

b) List and categorization of system stakeholders, users, and clients.

- **Stakeholders:**
  - Funding Entities: Organizations that provide the financial resources for the project.
  - Academic Community: Researchers and educational institutions.
  - Tech Companies: Entities that may use the system's capabilities to enhance their own products or services.
  - Machine Manufacturers: Companies that could in target the system into their customer service to provide automated support.

- **User:**
  - Remote Sensing Analysts: Professionals who analyse satellite imagery and could leverage the VQA system for enhanced interpretation.
  - GIS Specialists: Experts who work with geographic information systems and require detailed visual analysis.
  - Machine Learning Enthusiasts: Individuals interested in exploring AI technologies.
  - Data Analysts: Professionals who could use the system to derive insights from complex visual datasets.

- **Client:**
  - Arabic Language Speakers: Individuals who will directly use the system for inquiries in Arabic.
  - Frequent Travelers: Users who may require quick tech solutions.
  - People Facing Technical Difficulties with Household Machines: Users who seek assistance with home appliance issues.
  - Tech Companies, GIS and Remote Sensing Companies: Businesses that might use the system as part of their service offerings.
  - Real Estate and Construction Firms: Entities that could use the system for planning and development tasks.
  - Disaster Response Organizations: Groups that require rapid and accurate analysis of satellite images.

c) Challenges encountered [and lessons learned] during the requirements gathering, analysis, and prioritization phases.

The challenges encountered during the requirements gathering, analysis, and prioritization phases of the project were multifaceted and required careful consideration and mitigation strategies. Some of the key challenges and lessons learned include:

- Understanding User Needs: One of the primary challenges was accurately understanding the diverse needs of the target user groups, including remote sensing analysts, GIS specialists, machine learning enthusiasts, and data analysts. Lesson learned: Conducting thorough user research, surveys, and interviews helped uncover specific pain points and requirements of each user category.

- Integration of Diverse Datasets: Integrating and managing diverse datasets such as the Visual Genome and RSVQA datasets posed technical challenges related to data preprocessing, normalization, and compatibility. Lesson learned: Developing robust data processing pipelines and ensuring compatibility across datasets was crucial for seamless integration and system functionality.

- Stakeholder Alignment: Aligning the interests and expectations of diverse stakeholders, including funding entities, academic communities, tech companies, and machine manufacturers, required effective communication and stakeholder management. Lesson learned: Establishing clear communication channels, setting realistic expectations, and regularly updating stakeholders on project progress helped foster collaboration and mitigate conflicts.

3.2 System Requirements List:

a) Functional requirements

- login Function:

  - Aim: To login to the Question answering software
  - Input: Username and password
  - Output: navigates to the page to put your input question and image
  - Process: To navigate to main page

- Image and Question Submission:
    - Aim: To allow users to upload images and input questions
    - Input: An image file and a text-based question.
    - Output: The system accepts the image and question, then queues them for processing.
    - Process: Validation of image format and question text followed by submission to the VQA model.
- User Feedback:
    - Aim: To collect user feedback on the accuracy.
    - Input: User feedback through rating system.
    - Output: store the feedback.
    - Process: Adding every feedback given.
- Logout Function:
    - Aim: To logout from the system
    - Input: User action or command to logout
    - Output: System logout
    - Process: When the user initiates the logout process, the system clears the user's session.

b) Non-functional requirements:

- Reliability: The mean time between failures should be at Maximum 1 time per year.
- Recoverability: The system should recover after a breakdown within approximately 5 mins.
- Performance: Loading pages shouldn’t exceed 2 seconds.

4: System Design

4.1 System Archiucture

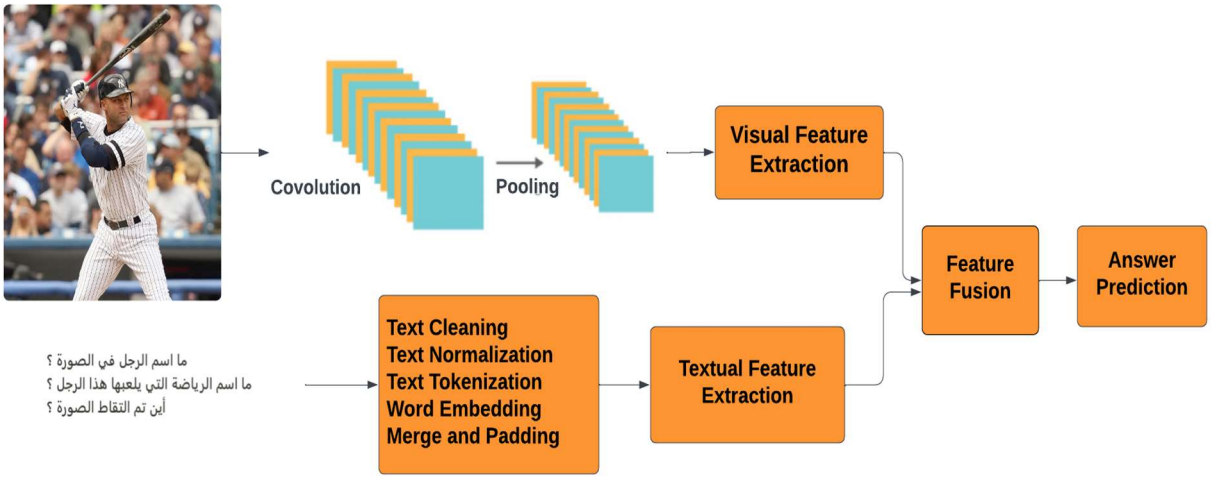


Fig.5. ANN System Architecture.

4.2. Algorithmic Components

- **Frame Extraction and Detection:** Use convolutional neural networks for object detection in images.
- **Visual Feature Extraction:** Processes segmented objects to create a detailed visual feature vector.
- **Arabic Language Processing:** Employs NLP techniques optimized for Arabic to analyze textual questions.
- **Data Fusion:** Integrates visual and language vectors for comprehensive contextual understanding.
- **Answer Generation:** Utilizes advanced AI models to generate contextually relevant answers in Arabic.
- **Optimization for Arabic:** Special emphasis on tailoring all processes for the Arabic language and context.

4.3 **Innovative Aspects of the Design:**

- **First Arabic VQA on RSVQA and Visual Genome Dataset:** Pioneering use of the RSVQA (Remote Sensing Visual Question Answering) dataset and Visual Genome Dataset, with a focus on Arabic language processing, a novel approach in the field.
- **Integration of Sentinel-2 Imagery:** Utilization of Sentinel-2 images (Retrieved from Google Earth Engine) combined with questions and answers derived from OpenStreetMap, creating a unique and rich database for analysis.
- **Extensive Database Coverage:** Work using 119,9508 images, more than 2.7 million Visual Question Answers.
- **Innovative Arabic Language Processing:** Specialized processing techniques to interpret and generate answers in Arabic, addressing a significant gap in current VQA research.

5. System Implementation

5.1. Hardware and Software Platforms

- Hardware: The system relies on high-performance GPUs to effectively train NLP models and handle image processing tasks. These GPUs enable parallel processing, enhancing user experience by providing quick and accurate responses.

- Operating System: Arabic-AssistQ is designed to be platform-independent and compatible with popular mobile operating systems like iOS and Android.
- Backend/Frontend Stacks: Flutter allows us to handle both the backend and frontend aspects of the application. It provides a rich set of UI components and efficient performance for creating attractive and responsive interfaces.

- AI Models: Arabic-AssistQ leverages the power of AI to provide accurate and context-aware assistance. We employ Large Language Models and natural language processing models to understand user queries, offer step-by-step guidance, and enable real-time translations.

- Augmented Reality (AR) Technology: To enrich the user experience, we integrate AR technology into Arabic-AssistQ. This involves utilizing frameworks like ARKit (for iOS) or ARCore (for Android) to enable real-time object recognition, text translation, and visual overlays.

5.2. Hardware and Software Development Tools and Languages:

- Development Tools: Development tools may include IDEs like PyCharm, Visual Studio Code, or IntelliJ IDEA for coding and debugging purposes. Additionally, tools for GPU-accelerated computing such as CUDA may be utilized.
- Languages: Programming languages such as Python are commonly used for implementing NLP models, image processing algorithms, and UI development. Additionally, languages like Dart may be used for developing the frontend UI components using frameworks like Flutter.

5.3. Modules/Components Acquired from External Sources:

- open-source libraries like NLTK, spaCy, and OpenCV provide foundational support for NLP and image processing tasks, the system will also leverage the PyTorch library for various AI-related tasks, including neural network development, training, and inference. PyTorch is renowned for its flexibility and efficiency in implementing deep learning models, making it

an ideal choice for tasks such as natural language processing (NLP) and image processing.

- Regarding licensed commercial or trial products: the system may consider utilizing OCR solutions from reputable providers or cloud-based services for high-quality text extraction from images. These tools can offer specialized features and enhanced accuracy, complementing the system's functionality in scenarios where precise optical character recognition is crucial.

- university and departmental resource libraries :play a vital role in providing access to cutting-edge research papers, datasets, and specialized libraries tailored to the system's AI and NLP requirements. Leveraging these resources enables the system to stay updated with the latest advancements in the field, driving innovation and improving performance in handling diverse linguistic and visual data.

5.4. Innovative Aspects of the Implementation:

- Integration of AI models: The system integrates cutting-edge AI models, including Large Language Models, for improved question understanding and tailored guidance.
- Use of Augmented Reality (AR) technology: Arabic-AssistQ incorporates AR technology to enhance the user experience by enabling real-time object recognition, text translation, and visual overlays.
- Seamless integration of backend and frontend: By using platforms like Flutter, the system ensures smooth communication between the backend AI models and the frontend UI, providing a seamless user experience across devices and operating systems.

6. Other Relevant Issues and Challenges.

The development of Arabic-AssistQ encountered several significant challenges, primarily revolving around the scarcity of readily available Arabic datasets. Despite the burgeoning interest in natural language processing (NLP) and computer vision, the limited availability of high-quality Arabic datasets posed a substantial impediment. In response, we amassed a collection of unstructured Arabic PDFs, aiming to fill this gap and facilitate the development and training of questioning and answering (Q&A) models tailored to Arabic language understanding. One of the most intricate tasks encountered was the process of extracting text from images within the PDF documents. Unlike digital text, where text recognition is relatively straightforward, extracting text from images proved to be a complex endeavour. This challenge necessitated the utilization of advanced optical character recognition (OCR) techniques to accurately extract, and process text embedded within images.

Furthermore, the dataset generation process presented its own set of challenges. Despite our efforts to create a comprehensive dataset, we faced efficiency issues with the tools used for data extraction. These challenges prompted the need for revisions and optimizations in our dataset generation pipeline to ensure the quality and relevance of the collected data.Moreover, the responses generated by the system often tended to be general rather than specific, posing additional challenges in providing precise and tailored guidance to user inquiries. Additionally, for lengthy PDF documents, extracting information from the end of the document posed logistical difficulties, requiring robust algorithms and techniques to navigate and extract relevant content efficiently.In addition to dataset-related challenges, the integration of satellite and visual genome datasets, originally in English, required translation into Arabic to align with the objectives of Arabic-AssistQ. This process of translation added complexity to the data integration phase, requiring meticulous attention to maintain accuracy and consistency across languages.

Despite these challenges, the team remained committed to overcoming obstacles and continuously refining the system to deliver an effective and reliable solution for Arabic language users. The lessons learned from addressing these challenges underscored the importance of robust data collection, efficient text extraction techniques, and precise translation processes in developing a successful Arabic-assistive system.

References

1. Ahmed, W., Bibin, P. A., & Anto, B. P. (2017). Question answering system based on neural networks. International Journal of Engineering Research, 6(3), 142-144.

2. Alsubhi, K., Jamal, A., & Alhothali, A. (2021). Pre-trained transformer-based approach for Arabic question answering: A comparative study. arXiv preprint arXiv:2111.05671.

3. Chen, C. H., Wu, C. L., Lo, C. C., & Hwang, F. J. (2017). An augmented reality question answering system based on ensemble neural networks. IEEE Access, 5, 17425-17435.

4. Das, A., Datta, S., Gkioxari, G., Lee, S., Parikh, D., & Batra, D. (2018). Embodied question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-10).

5. Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J.

M., ... & Batra, D. (2017). Visual dialog. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 326-335).

6. Kamel, S. M., Hassan, S. I., & Elrefaei, L. (2023). VAQA: Visual Arabic Question Answering. Arabian Journal for Science and Engineering, 1-21.

7. Mozannar, H., Hajal, K. E., Maamary, E., & Hajj, H. (2019). Neural Arabic question answering. arXiv preprint arXiv:1906.05394.

8. Tan, S., Ge, M., Guo, D., Liu, H., & Sun, F. (2023). Knowledgebased embodied question answering. IEEE Transactions on Pattern Analysis and Machine Intelligence.

9. Wong, B., Chen, J., Wu, Y., Lei, S. W., Mao, D., Gao, D., & Shou, M. Z. (2022, October). Assistq: Affordance-centric question-driven task completion for egocentric assistant. In *European Conference on Computer Vision* (pp. 485-501).

10. Yu, L., Chen, X., Gkioxari, G., Bansal, M., Berg, T. L., & Batra, D. (2019). Multi-target embodied question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6309-6318).