7 Methodology

This section specifies the proposed system design, the datasets that will be used to evaluate the system performance, the evaluation metrics, the experiments that will be conducted, and the limitations.

7.1 The Proposed Arabic VQA System Design

The proposed modular Arabic VQA system consists of the two modules shown in Figure 4: the IC Module and the Answer Generation Module.

- **The IC Module:** This module consists of two pre-trained, Arabic IC models, which will allow the extraction of the image's contextual information to be maximised. There are many pre-trained Arabic IC models [2, 49, 69, 19], but the Arabic IC models that will be adopted for the proposed system include a pre-trained BiT [19] and a VLM called Violet [49]. They were chosen because they employ advanced deep learning methods, such as transformers and attention, and they demonstrate good Arabic IC performance. An image will be input into these IC models, after which they will generate a set of captions that represent the image's context. The generated captions will be filtered to remove captions that are exact substrings of others. The best N captions will then be selected to exclude noisy captions (this selection process is further discussed in Section 8.4.2). Different N values will be used during the evaluation. Finally, the image's extracted context will be forwarded to the answer generation module in the form of a text prompt.
- **The Answer Generation Module:** This module consists of an Arabic LLM. Many pre-trained Arabic LLMs are currently available, such [7, 1, 9, 31, 8]. The system proposed herein

will adopt AraGPT-2 [9] because it meets the following criteria:

1. the ability to perform Arabic question-answering with zero-shot settings [9]; and
2. the ability to leverage its implicit knowledge or commonsense reasoning [40, 22, 38, 41].

This model will receive an input text prompt that consists of a prompt head, the contextual information extracted from the image, and the question all in Arabic. Figure 5 shows the input text prompt formulation from left to right. The model will be provided with the prompt head 'Answer the question from the following context.', then the phrase 'Context:' and a sequence of N captions, which will be followed by the phrase 'Question:' and the question, all in Arabic language.

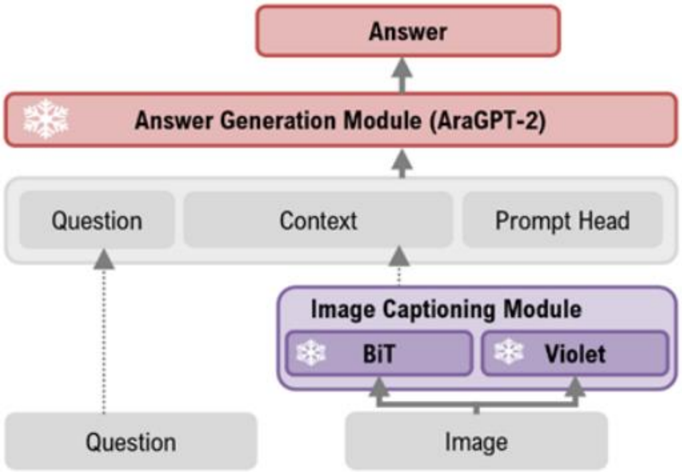

Figure 4: The Architecture of the Proposed Modular Arabic VQA System

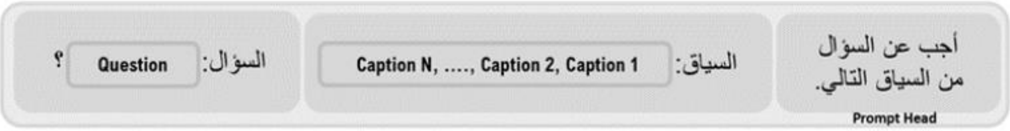

Figure 5: The Formulation of the Inputted Prompt Text

A single image and question in Arabic about the image's content will be input into the system, and an answer in standard Arabic will be generated using LLM's commonsense reasoning and the image's contextual information. Figure 6 below gives two examples to explain the general nature of inputs and outputs each module in the system will receive or generate.

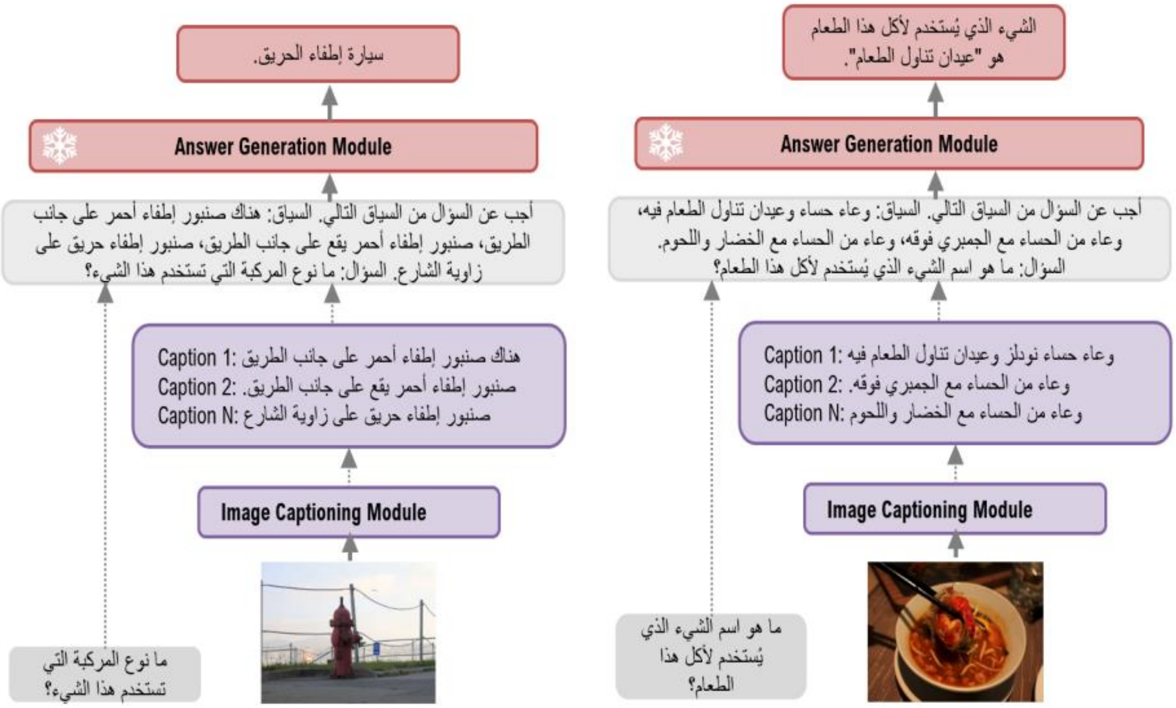
سيارة إطفاء الحريق.
Answer Generation Module
أجب عن السؤال من السياق التالي. السياق: هناك صنبور إطفاء أحمر على جانب
الطريق، صنبور إطفاء أحمر يقع على جانب الطريق، صنبور إطفاء حريق على
زاوية الشارع. السؤال: ما نوع المركبة التي تستخدم هذا الشيء؟
Caption 1: هناك صنبور إطفاء أحمر على جانب الطريق
Caption 2: صنبور إطفاء أحمر يقع على جانب الطريق.
Caption N: صنبور إطفاء حريق على زاوية الشارع
Image Captioning Module
ما نوع المركبة التي
تستخدم هذا الشيء؟
الشيء الذي يُستخدم لأكل هذا الطعام
هو "عيدان تناول الطعام".
Answer Generation Module
أجب عن السؤال من السياق التالي. السياق: وعاء حساء وعيدان تناول الطعام فيه،
وعاء من الحساء مع الجمبري فوقه، وعاء من الحساء مع الخضار واللحوم.
السؤال: ما هو اسم الشيء الذي يُستخدم لأكل هذا الطعام؟
Caption 1: وعاء حساء نودلز وعيدان تناول الطعام فيه
Caption 2: وعاء من الحساء مع الجمبري فوقه.
Caption N: وعاء من الحساء مع الخضار واللحوم
Image Captioning Module
ما هو اسم الشيء الذي
يُستخدم لأكل هذا
الطعام؟

Figure 6: Examples of expected modules' inputs and outputs

7.2 Datasets

Due to the scarcity of high-quality Arabic VQA data as well as the time-consuming and costly process of manually annotating and creating a new dataset, we will generate the evaluation datasets for our system by translating widely used EVQA datasets using Google Translation API. Two datasets will be used.

- VQAv2 [20]: This dataset consists of one million questions about 265,016 images.
- OK-VQA [46]: This dataset comprises more than 14,000 questions that require commonsense reasoning or external knowledge to answer them. It covers a wide range of knowledge categories.

Following the translation, we will filter out poorly translated samples from the dataset by measuring the similarities between the sentences of the original text and the translated text using the BLUE-N score [52]. The similarity threshold score will be selected based on observations.

7.3 Evaluation

We will evaluate the Arabic VQA system's performance using several metrics.

- **Accuracy**: This is defined as the ratio of correctly predicted answers to the total number of ground-truth answers for a set of questions. The accuracy computation involves comparing the system's answer to multiple ground-truth answers for each question. It is computed using the following equation:

  $$Accuracy = \min\left(\frac{NumOfMatchingAns}{3}, 1\right) \tag{1}$$

  Where the **NumOfMatchingAns** refers to the number of ground truth answers that match the system answer for a question. After that, the overall accuracy of the system is the average of individual question accuracies. This metric was proposed by [20] to ensure robust inter-human variability in phrasing the answers.

- **BLUE-N Score** [52]: This variable denotes the bilingual evaluation understudy. It was originally proposed as an MT metric, but due to it being challenging to evaluate generative answers, particularly if they are long, many VQAs use it to determine the similarities between the actual and generated answers [25] based on n-gram precision (n-grams are contiguous sequences of n-words). The precision will be calculated for n-grams of varying lengths, and the geometric mean of these precisions will be used to compute the BLUE-N score. The formula for the BLUE-N score is as follows:

  $$BLUE - N = BP \exp\left(\sum_{n=1}^{N} \frac{1}{N} \log precision_n\right) \tag{2}$$

  where **N** is the maximum order of n-grams considered and $precision_n$ is the precision of n-grams, calculated as the ratio between the number of n-grams in the generated answer included in the actual answer to the total number of n-grams in the generated answer.

$$precision_n = \frac{NumberOfMatchingNGrams}{TotalNumberOfNGramsInGeneratedAnswer} \tag{3}$$

where **BP** denotes the Brevity Penalty, which is the penalty term employed to address the issue of shorter translations being favoured that is calculated as follows:

$$BP = \min\left(1, \frac{lengthOfGeneratedAnswer}{lengthOfActualAanswer}\right) \tag{4}$$

The BLUE-N score ranges from 0 to 1, with values closer to 1 indicating greater similarities.

7.4 Experiments

Two types of experiments will be conducted to evaluate the system performance by manipulating the image captioning methods and manipulating the image caption selection methods.

7.4.1 Image Captioning

An ablation study will be conducted on the IC Module to identify how to best represent images in the textual format for AraGPT-2. Specifically, several methods will be compared.

- **Blinded:** The blinded system settings use an empty string to represent the image's context, which will be used to indicate the Arabic VQA's performance without the image's contextual information. It will also reflect the LLM's implicit knowledge.
- **BiT [19] Captions:** The image's context will only be represented by the Bidirectional Transformer [19] model.
- **Violet Captions:** The image's context will only be represented by the Violet [49] model.
- **Combination of all captions:** The image's context will be represented by captions generated from all the models in the IC module.

7.4.2 Image Caption Selection

We will conduct an ablation study on the caption selection methods below according to the BLUE-N scores obtained.

- **Random:** This will determine the system's performance when the captions are randomly selected.
- **High similarity to the question:** This will indicate the system's performance when the captions are selected based on their similarity to the question.
- **Low similarity to the question:** This will denote the system's worsened performance when noisy captions are selected.
- **High similarity to the questions and answers:** This will specify the system's performance when the selected captions have high similarity to the answers. It will be used as an upper bound.

7.5 Limitations

Both of the modular Arabic VQA system's features and limitations are the result of the zero-training modular system paradigm. While our system leverages the pre-trained models' strengths, it will also inherit their biases and limitations. The system will take advantage of the efficiency of zero training, but it will also incur additional computation costs because of the additional stages involved. In brief, the limitations are as follows:

1. The proposed system will inherit the biases of the pre-trained models; and
2. The proposed system's performance will be limited by the adopted pre-trained models' performance.
3. Limited accessibility may be encountered for the chosen Arabic IC models. Alternative models meeting the same criteria will be considered to overcome this limitation.

9 Expected Findings

After implementing the proposed modular Arabic VQA system and achieving all the required objectives, the following findings are expected.

1. A new Arabic VQA system.

   We will develop an Arabic VQA system capable of answering an open-ended question about a single image in Arabic.

2. Performance evaluation of the Arabic VQA system.

   With the help of BLUE-N score to measure the Arabic VQA performance, we will manipulate the Arabic IC models and caption selection methods to analyze the best design that maximizes the VQA performance.

3. Impact of Arabic IC on Arabic VQA.

   We will investigate the impact of different Arabic IC models on the performance of the modular Arabic VQA system

4. Impact of caption selection on Arabic VQA.

   We will investigate the impact of different caption selection methods on the performance of the modular Arabic VQA system.

5. Impact of Arabic LLM's implicit knowledge on Arabic VQA.

   We will investigate the impact of AraGPT-2's implicit knowledge on the performance of the modular Arabic VQA system.

6. Insights into the challenges and opportunities of adopting modular paradigm in Arabic VQA.