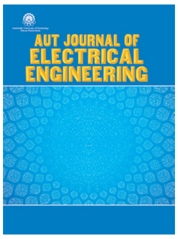

AUT Journal of Electrical Engineering

AUT J. Electr. Eng., 56(2) (2024) 191-202
DOI: 10.22060/eej.2023.22605.5552

Envisioning Answers: Unleashing Deep Learning for Visual Question Answering in Artistic Images

Erfan Zolghadriha[1] , Kazim Fouladi-Ghaleh[2]* , Pouya Ardehkhani[1]

[1] Deep Learning Research Lab, Department of Computer Engineering, Faculty of Engineering, College of Farabi, University of Tehran, Iran
[2] Department of Computer Engineering, Faculty of Engineering, College of Farabi, University of Tehran, Iran

ABSTRACT: In specialized fields, the accurate answering of visual questions is crucial for practical applications, and this study focuses on improving a visual question-answering model for artistic images by utilizing a dataset with both visual and knowledge-based questions. The approach involves employing a pre-trained BERT model to understand query nature and using the iQAN model with MLB and MUTAN mechanisms for visual queries, along with an XLNet-based model for knowledge-based information. The results demonstrate a 78.92% accuracy for visual questions, 47.71% for knowledge-based questions, and an overall accuracy of 55.88% by combining both branches. Additionally, the research explores the impact of parameters like the number of glances and activation functions on the model's performance.

Review History:

Received: Aug. 08, 2023
Revised: Sep. 09, 2023
Accepted: Sep. 26, 2023
Available Online: Mar. 01, 2024

Keywords:

Art Pictures
Visual Question Answering (VQA)
Natural Language Processing (NLP)
Computer Vision
Attention

1- Introduction

The concept of answering visual questions is a recent and fascinating area of study in deep learning, which extends the idea of machine understanding. Visual Question Answering (VQA) is an interdisciplinary task in artificial intelligence that combines advanced techniques in computer vision and natural language processing. The goal is to develop a system capable of answering questions about images by comprehending their meaning and semantics. Previous research in this field has predominantly focused on utilizing deep learning architectures and various learning algorithms from computer vision, object recognition, and natural language processing domains. These approaches aim to achieve the objective of answering questions based on information derived from the images. However, most studies have primarily been conducted on datasets without specific content categories, where images are organized into different categories such as cars, people, places, animals, etc.

Recently, there has been an increasing interest in answering visual questions in more specialized domains like medicine and art. These domains require processing images and questions that are specific to their respective fields, thereby enhancing the applicability of VQA systems in knowledge-dependent areas. Art and computer vision are inherently linked due to the visual components present in artistic elements. Digitizing artworks for preservation and restoration purposes is a fundamental step within this field. Extensive research has been conducted in the realm of computer vision pertaining to works of art, including tasks such as style and author identification [1,2], image categorization [3-7], and image retrieval [8-10].

In 2018, the SemArt dataset was introduced for the semantic understanding of art, containing paintings and related comments serving as metadata for the artwork [11]. Building upon SemArt, the AQUA dataset was created in 2020 [12]. AQUA aimed to facilitate the task of visual query answering in art images. Unlike previous datasets, AQUA focused not only on the visual content of the paintings but also incorporated opinions associated with them. The question-answer pairs in this dataset contain visual states and demand knowledge beyond the images themselves. In the same article introducing the AQUA dataset [12], a preliminary model named VIKING was presented as the initial attempt to address the problem of answering visual questions in art images. The VIKING model comprises three main components.

The "Modality Selector" section specifies the nature of the questions and divides them into two visual and knowledge-based categories. In the visual branch, the VIKING model predicts answers to the given questions that can be answered solely from the images using the basic iQAN model [13]. The iQAN model is a dual model that takes a question or answer as input and generates its counterpart as output. Some

*Corresponding author's email: kfouladi@ut.ac.ir

visual questions in the AQUA dataset are also generated using this model. Questions classified as knowledge-based are categorized under answering questions based on external knowledge. This branch consists of two parts. Firstly, a two-stage process retrieves external knowledge related to each question and ranks the relevance of opinions with respect to the question, resulting in a subset of ten related opinions for each question. Subsequently, for knowledge-based questions, an answer is predicted using an XLNet model [14]. In this paper, we implement and enhance the performance of the VIKING model using the AQUA dataset. To achieve this, we replace the MUTAN attention mechanism [15], employed in the basic model, with the MLB attention mechanism [16]. We combine the MLB attention mechanism with the MUTAN fusion mechanism in the iQAN model and the slightly modified MUTAN Fusion model so that it works well with the MLB attention mechanism. Our primary goal in improving VIKING is to strike a balance between reducing model complexity and enhancing accuracy, with a specific focus on Visual Question-Answering systems. This involves streamlining the model architecture while prioritizing its performance to ensure more efficient and effective answers to visual queries. In Section 3, we provide a more comprehensive explanation of our proposed method.

In the following sections, we will delve into our proposed methodology for addressing the challenges posed by visual question-answering. Section 2 provides an overview of previous studies in this domain, highlighting the existing research landscape. In Section 3, we introduce our novel approach, which consists of distinct branches: Section 3.1 explains our choice of approach, while Section 3.2 delves into the answer branch dedicated to visual questions, and Section 3.3 focuses on the branch tailored to answering questions based on external knowledge. Subsequently, in Section 4, we present our experimental results. Section 4.1, we delve into the specifics of our dataset, the AQUA dataset. Section 4.2, discusses the outcomes of modality selection, while Sections 4.3 and 4.4 present results pertaining to the visual and external knowledge branches, respectively. In Section 4.5, we provide a comprehensive summary of our final results, and in Section 4.6, we perform a comparative analysis of our proposed method against existing approaches. In section 5, we present questions that can be answered in future studies and works. Finally, we conclude our paper in Section 6 by summarizing our key findings and discussing their broader implications.

2- Previous Studies

A lot of attention has been paid recently to deep learning-based visual question-answering schemes. In the following, we will examine several of the models presented to solve the problem of answering image questions.

Attention-based VQA: In 2015, Chen et al introduced a tunable convolutional neural network model based on the attention process [17]. Attention-based Visual Question Answering (VQA) models integrate visual attention mechanisms, allowing them to focus on specific regions of an image while answering questions. These models dynamically weigh different parts of an image to provide accurate answers to questions based on both image and textual inputs.

Fact-based VQA: This model is based on discarding questions that we need real knowledge to answer (Wang et al. 2016 [18]). Fact-based VQA models excel at providing answers grounded in factual information. They prioritize factual accuracy and rely on knowledge bases and structured data sources to validate their responses, making them valuable for tasks requiring precise and verifiable answers.

Focus Regions for VQA: This method is based on learning to select areas of the image that are related to the questions and answers provided (Shih et al. 2016 [19]). Focus Regions models enhance VQA by explicitly identifying and emphasizing critical regions within an image that are most relevant to the posed question, thus improving answer quality and interpretability.

Focused Dynamic Attention (FDA): This model for better alignment of image content representation with questions was presented in 2016 by Ilievski et al [20]. FDA models extend traditional attention mechanisms by dynamically adjusting attention weights based on the context of the question. This adaptability helps these models better capture intricate relationships between image and text inputs.

Dual attention network for VQA: In this work, instead of applying the attention process only to the images, it was applied to both the images and the questions. This work was introduced in 2017 by Xu and Saenko [21]. Dual Attention Networks introduce two separate attention mechanisms—one for visual features and one for textual features. This dual attention approach improves VQA performance by considering both modalities more effectively.

Structured attention for VQA: In 2017, Zhu et al. proposed a visual attention model based on a multivariate distribution on a grid-structured conditional random field in image regions [22]. Structured Attention models incorporate structured representations into VQA, allowing them to handle complex questions by breaking them down into more manageable sub-tasks. This structured approach enhances interpretability and reasoning capabilities.

VQA-E: Most VQA algorithms focus on improving the accuracy of response prediction, but they ignore the description. Qing Li [23] believed that the description of the response is as important as or even more important than the response. Because this makes the process of understanding questions and answers easier and gives more information to blind people. For this reason, Qing Li introduced the VQA model (VQA-E with description). In this work, a data set called VQA-E was extracted from the second version of the VQA data set, and then the multi-task model of VQA-E was introduced on the provided data set. VQA-E models focus on addressing questions with an emotional or affective dimension. They analyze both visual and textual content to provide answers that reflect the emotional context of the input, making them suitable for applications involving sentiment analysis.

Differential networks: This model uses the differences between forward propagation steps to reduce noise and learn

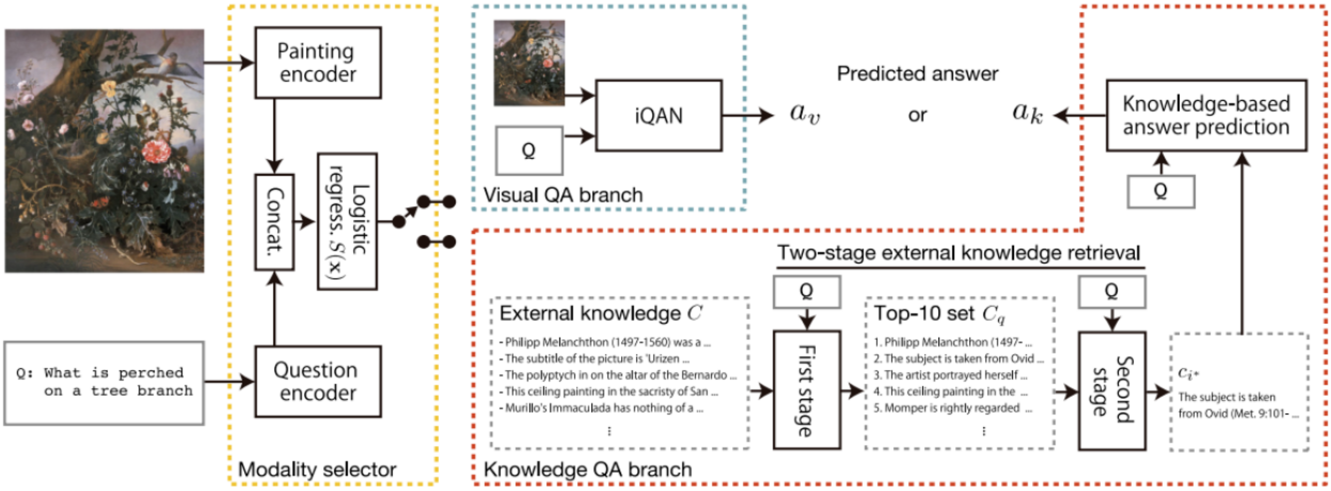
Painting encoder
Concat.
Logistic regress. S(x)
Q: What is perched on a tree branch
Question encoder
Modality selector
iQAN
Q
Visual QA branch
Predicted answer
a_v
or
a_k
Knowledge-based answer prediction
Q
Two-stage external knowledge retrieval
External knowledge C
- Philipp Melanchthon (1497-1560) was a ...
- The subtitle of the picture is 'Urizen ...
- The polyptych in on the altar of the Bernardo ...
- This ceiling painting in the sacristy of San ...
- Murillo's Immaculada has nothing of a ...
Q
First stage
Top-10 set C_q
1. Philipp Melanchthon (1497- ...
2. The subject is taken from Ovid ...
3. The artist portrayed herself ...
4. This ceiling painting in the ...
5. Momper is rightly regarded ...
Q
Second stage
c_{i^*}
The subject is taken from Ovid (Met. 9:101- ...
Knowledge QA branch

Fig. 1. The proposed method

the interdependence between features [24]. Image features are extracted using Faster-RCNN [25]. Differential modules [26] are used to modify features in text and images. GRU is used to extract query features. Finally, it is combined with an attention module to classify responses. Differential Networks introduce specialized modules that quantify the differences and relationships between visual and textual information, enabling more nuanced reasoning in VQA tasks.

Joint Embedding VQA Model: Ma et al. introduced this approach in 2021 [26] and concluded that dynamic word vectors outperform static word vectors in the VQA task. Joint Embedding models aim to bridge the gap between vision and language by learning a shared embedding space for both modalities. This shared space facilitates the seamless integration of image and text information for more accurate VQA.

The paper titled "Flamingo: a Visual Language Model for Few-Shot Learning" authored by a team of researchers introduces the Flamingo model in the year 2022 [27]. Flamingo addresses the challenge of rapidly adapting models to novel tasks with limited annotated examples in the field of multimodal machine learning. The model incorporates several key architectural innovations, including bridging pretrained vision-only and language-only models, handling sequences of mixed visual and textual data, and seamlessly processing images or videos as inputs. This flexibility enables Flamingo to be trained on large-scale multimodal web corpora, making it capable of in-context few-shot learning. The authors conducted a comprehensive evaluation of the Flamingo model, demonstrating its efficacy across various image and video tasks, including visual question-answering, captioning, and multiple-choice visual question-answering. Flamingo consistently outperforms models fine-tuned on significantly larger amounts of task-specific data, establishing itself as a state-of-the-art solution for few-shot learning in multimodal applications.

3- Proposed Method

In this work, the basic model presented in [12] has been used to solve the problem of answering visual questions in artistic images and applying some changes to improve the performance of this model. This model has three general parts; The modality selector section, the visual branch, and the knowledge-based branch, which we will introduce in the next sections.

3- 1- Choice of Nature (Modality Selector)

In this section, we will present the approach employed in the modality selector section for encoding and extracting feature vectors from the questions and drawings in the dataset. These features are then combined to determine the nature of the questions. To accomplish this, the model utilizes a pre-trained BERT model [28] as a question encoder to extract textual features. Specifically, the model employs the BERT-Large, Uncased model as its base model, generating a 1024-dimensional vector q to encode the question. Additionally, a pre-trained ResNet-152 architecture has been leveraged [29] to encode the paintings into a 2048-dimensional vector v, enabling us to extract the painting features. Next, a modality selector (S) engages to categorize a question q based on the combined vectors v and q into one of two categories: visual questions or questions requiring external knowledge to answer the questions. To achieve this, feature vectors v and q emerge, creating the vector X. Then the vector X will be

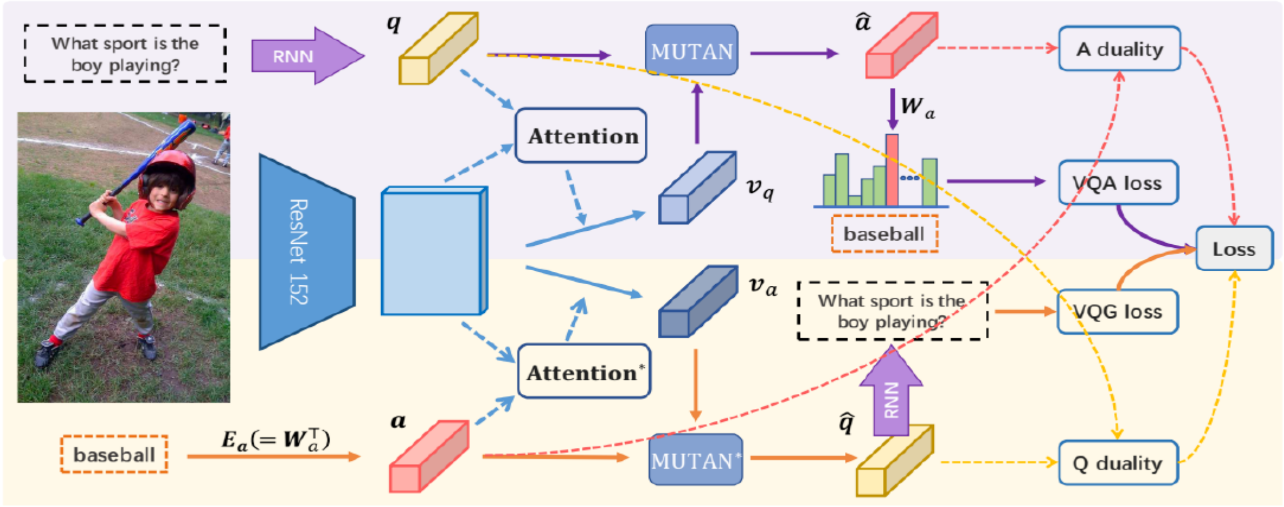

Fig. 2. iQAN architecture

applied to a logistic regression model using Equation (1) to perform the modality selection process.

$$S(X)=\frac{1}{1+e^{-\left(w_s^T X-b_s\right)}} \qquad (1)$$

In Equation (1) [30], the parameters w_s and b_s are numerical vectors that can be trained. A question q is given to the answer branch of the visual question if the value of $S(X)>0.5$. Otherwise, the query q is submitted to the knowledge-based branch.

3- 2- Answer Branch to the Visual Question

Visual questions can be answered based on the associated drawing alone, without any external knowledge. For these types of questions, the work is reduced to answering the visual question on the drawings. In this research, in the iQAN model, we use the MLB attention mechanism module instead of the MUTAN attention mechanism module, along with the MUTAN fusion mechanism as the answer branch to our visual question. The attention module has the task of increasing the focus on the points of the input images that are more relevant to the question. The fusion module is also responsible for combining textual and visual features to reach response vectors. For this purpose, we separately train the iQAN model on the training partition of the AQUA dataset. This branch produces a predicted answer a_v which is from the answer vocabulary A consisting of 5000 common words in the educational division.

The iQAN model or reversible question-answering network is an end-to-end integrated model. In this model, using the reversible bilinear fusion module and the parameter sharing scheme, both the task of answering the visual question and the task of generating the visual question can be performed simultaneously. By jointly training these two tasks with dual regulators (dual training), the model will better understand the interactions between images, questions, and answers. After training, iQAN can take a question or answer as input and generate its counterpart in output.

In this research, in the visual question-answering part, according to the question, an RNN is used to obtain the embedded feature q, and on the other hand, a CNN is used to convert the input image into a feature map. An attention module based on the MLB model [16] is used to create a query-aware image feature vector v_q. Then, using a MUTAN [15] fusion module, other response feature vectors $\hat{a}$ are obtained by merging v_q and q. Finally, a linear classifier w_a predicts the response. The visual question-answering part in the basic iQAN model is based on one of the advanced visual question-answering models, MUTAN. In this research, we used the implemented attention mechanism of the MLB visual question-answer model in combination with the MUTAN fusion mechanism, and for this reason, we modified the iQAN model.

Also, one of the superiorities of the MLB (Multimodal Low-Rank Bilinear) attention mechanism over the MUTAN (MUtual TANdem) attention mechanism lies in its ability to efficiently capture complex intermodal interactions between different modalities in multimodal tasks, such as image and text processing. MLB employs a low-rank bilinear pooling approach, which significantly reduces the computational complexity compared to MUTAN's quadratic interactions. This computational efficiency makes MLB more suitable for real-world applications, particularly when dealing with large-scale datasets and resource-constrained environments, without sacrificing the modeling power necessary to capture intricate cross-modal relationships. Consequently, MLB offers a compelling trade-off between performance and computational cost, making it a preferred choice in many multimodal tasks.

3- 3- Branch of Answering Questions Based on External Knowledge

The questions categorized as knowledge-based fall under the category of answering questions that rely on external information. To identify the most relevant comment in a given context, a two-step strategy is employed. Initially, the comments in the context are ranked based on their relevance to the question using TF-IDF in the base model. From this ranking, a subset called C_q is generated, consisting of the top 10 comments that are deemed most relevant to the question among all the comments in the context.

To determine the score between the TF-IDF vectors of the question ($\hat{q}$) and the comment ($\widehat{c_i}$), Equation (2) [31] is utilized. The calculation of the score is as follows:

$$S_i = \frac{\hat{q}^T \hat{c}_i}{\left(\|\hat{q}\|\|\hat{c}_i\|\right)} \tag{2}$$

where i is an index or identifier for different comments in a collection and q is a vector, representing a transformed or processed version of a question. The collection C_q comprises the top 10 elements C_i with the highest values of S_i . To enhance the accuracy of the ranking process for both the query (q) and the elements C_i , various techniques are employed. NLTK is utilized for stopping word deletion and word rooting, while TF-IDF is applied in the n-grams mode with a window size of 3.

Moving on to the next step, the objective is to identify the opinions relevant to the query. In this case, the ranking of member C is carried out. To accomplish this, a binomial classification approach is employed, utilizing a BERT model based on Bert-base-uncased. This BERT model predicts the probability of the extent to which a given C_i is related to the query q. The concatenation of q and each member C_i from C is achieved using the [SEP] token, followed by feeding it into the pre-trained BERT model. The BERT model leverages special symbols, namely [CLS] and [SEP], to accurately comprehend the input. The resulting output, denoted as O_i associated with the [CLS] symbol, is then passed through a logistic regression model obtained via Equation (3) [28]:

$$R(o_i) = \frac{1}{1+e^{-\left(w_r^T o_i - b_r\right)}} \tag{3}$$

In equation (3), the variables w_r and b_r represent vectors that can be adjusted through training. The purpose of the model is to classify whether q (a question) and c_i (a candidate answer) are related or not, acting as a binary classifier. However, during the inference stage, the output $R(o_i)$ of the model is treated as the score assigned to c_i . In the basic model, the candidate's answer c_{i^*} with the highest score, determined by $i^* = \arg\max_i R\left(o_i\right)$, is selected to answer the question.

To predict answers to questions using external knowledge, the basic model utilizes XLNet [14]. XLNet is an extension of the Transformer-XL model [32] and operates as an autoregressive transformer. It combines the strengths of autoregressive language modeling and autocoding while aiming to overcome their limitations. Pre-training of the XLNet model is accomplished using a self-reversal method that enables learning from both forward and backward contexts. The training maximizes the expected likelihood across all possible permutations of the input sequence's factorization order.

In the query stage, q (the question) and c_{i^*} (the selected candidate's answer) are concatenated with the special token [SEP], forming a single input sequence. This combined sequence is then provided to the XLNet model. XLNet predicts the start and end positions of the response, indicating the answer's span within the concatenated sequence. The words located between the predicted start and end positions are extracted as the final answer, denoted as a_k .

The model employs a pre-trained XLNet, which is specifically fine-tuned on question-answer pairs using the knowledge from the AQUA dataset [12]. This pre-training enhances the model's ability to generate accurate responses to questions based on the information contained in the dataset.

4- Results

In this section, we review the results of the tests obtained in different branches. The performance of our work is measured by exact match (EM), i.e., the percentage of predictions that exactly match the reference data.

4- 1- Dataset

In this study, we utilize the AQUA dataset to address visual inquiries pertaining to artistic images. To ensure originality and provide further details, we extend upon the given context. The AQUA dataset is derived from the SemArt dataset [11], which comprises drawings along with associated textual comments serving as a source of knowledge. These comments play a significant role in our research. Our objective is to showcase the capabilities of artificial intelligence technologies in comprehending paintings. To achieve this, we employ question and answer pairs that encompass two visual states and rely on external knowledge. We adopt specific methods corresponding to these states to generate the questions.

Within the dataset, two distinct approaches are employed for generating visual questions. The first approach involves utilizing iQAN, which is trained on the second version of the VQA dataset [33]. iQAN takes an image and an answer word as inputs and employs a neural network model to generate a

Table 1. Statistical details of the AQUA dataset. Note that the length of questions and responses are average tokens generated in a partition of our dataset.

Number	Train	Validation	Test
Number of QA pairs	69812	5124	4912
Visual	29568	1507	1270
Knowledge	40244	3617	3642
Question length	8.82	9.21	9.41
Visual	6.53	6.50	6.51
Knowledge	10.50	10.33	10.43
Length of responses	3.13	3.68	3.85
Visual	1.00	1.00	1.00
Knowledge	4.69	4.79	4.85

Table 2. Error matrix of nature selector

Label	Prediction	
	Visual	Knowledge
Visual	1269	1
Knowledge	17	3625

query. The second approach involves leveraging Pythia [34] to generate a title for each drawing. Subsequently, using the rule-based Technique for Question Generation (TQG) [35], each title is transformed into a set of paired questions and answers.

To generate questions that necessitate knowledge about art for their answers, the AQUA dataset employs TQG methods. Multiple TQG strategies, such as rule-based and neural approaches [35], have been tested on this dataset. For a comprehensive overview of the AQUA dataset, please refer to Table 1, which presents the dataset statistics.

4- 2- Results of Modality Selection

In a manner similar to the basic model [13], our study achieved a remarkable accuracy of 99.6% in distinguishing between visual questions and knowledge-based questions during the nature selection phase. The classifier was able to differentiate them easily due to the diverse methods employed in creating the visual and knowledge-based questions. The selector-error matrix, presented in Table 2, provides further insights into the classification accuracy of the question types.

4- 3- The Results of the Branch Based on Visual Knowledge

Within the training domain of visual question-answering, our research focused on providing accurate responses to a total of 1286 questions. We successfully answered 1015 questions correctly, resulting in an overall accuracy of 78.92% in the visual branch of question-answering. To evaluate the performance, Figure 3 depicts the accuracy graph based on Acc@k (where k represents the number of selected answers), specifically on the validation set. For instance, Acc@5 signifies the presence of the correct answer within the top five selected options. Additionally, Figure 4 illustrates the loss graph in the validation set. It is important to note that the reported accuracy represents the highest achieved accuracy by our proposed model.

The striking similarity between Acc@5 and Acc@10, with both achieving similar accuracy, can be attributed to several factors. Firstly, the AQUA dataset, a derivative of SemART, may exhibit a certain level of inherent redundancy

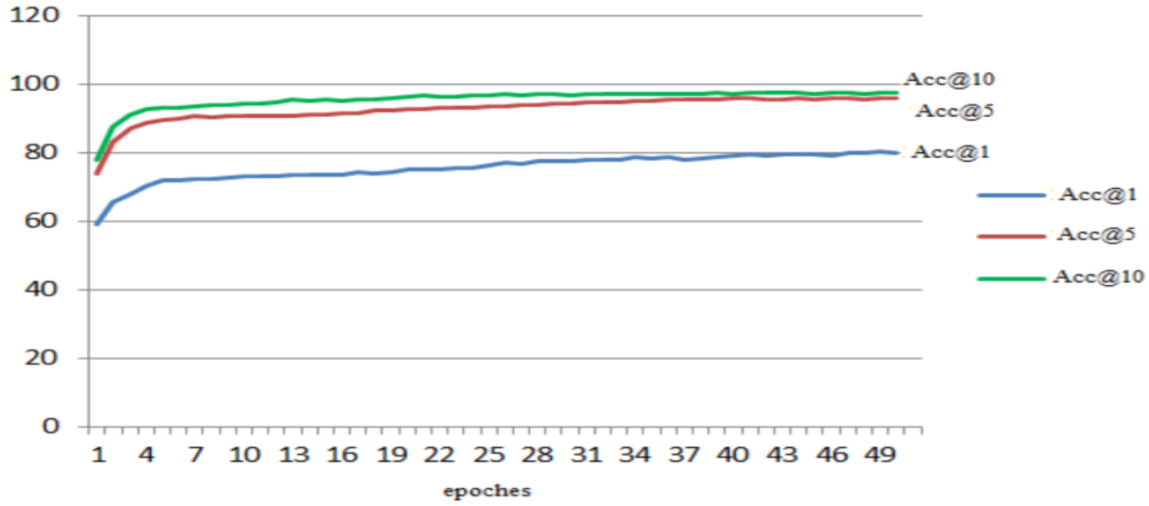

Fig. 3. Accuracy Graph

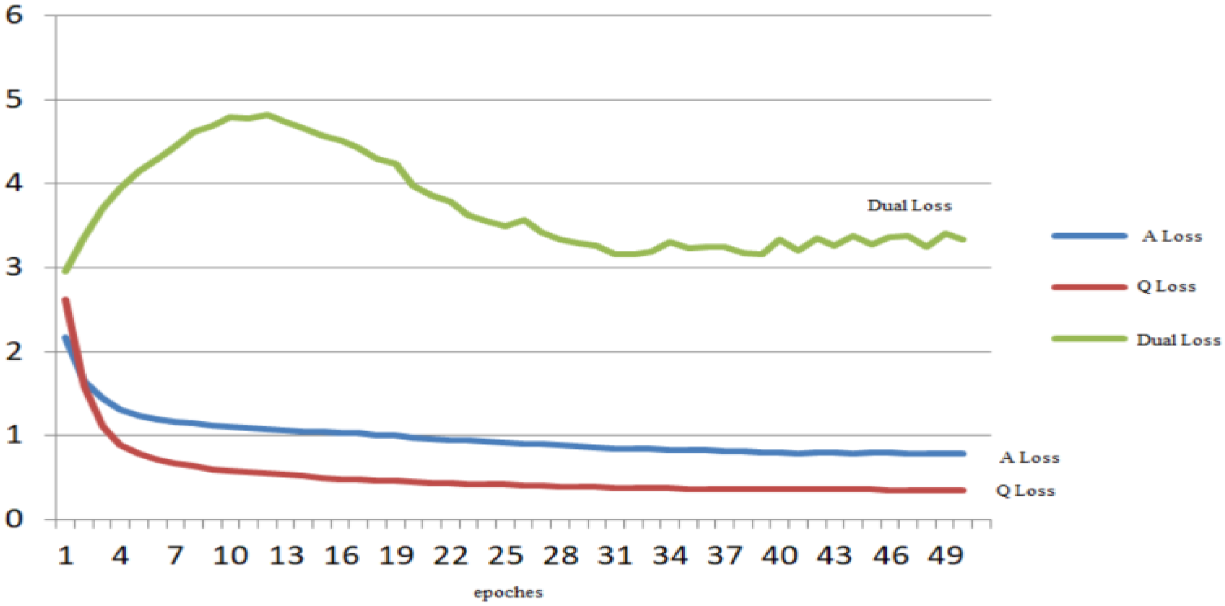

Fig. 4. Graph of the amount of loss on the validation division

or predictability within its content, which allows the model to make confident predictions even beyond the top 5 candidate answers. Secondly, the choice of the iQAN model for question answering, coupled with the utilization of a ReLU activation function and the use of four glances during inference, may collectively contribute to consistent performance across these two evaluation metrics. The ReLU activation function, for instance, is known for its ability to capture non-linear patterns effectively, while the incorporation of four glances allows the model to gather more contextual information, potentially enhancing its ranking capabilities. Consequently, the minor discrepancy between Acc@5 and Acc@10 may primarily stem from the diminishing returns of additional candidates beyond the top 5, given the relatively high accuracy achieved within the initial candidate set. Further exploration and analysis of model behavior under different settings may

Table 3. The results of the branch based on foreign knowledge

Comments retrieval stage	Attending the first top comment	77.16%
	Being in the top 5 comments	88.08%
	Being in the top 10 comments	90.95%
Response prediction stage	Number of correct answers	1730
	The number of knowledge questions	3626
	Final accuracy	47.71%
	f1 Score	58.522

provide deeper insights into this intriguing pattern.

The optimal accuracy was attained through the use of the tanh activation function [36] and employing 2 glances at the image during the attention mechanism training. This combination led to the highest accuracy in answering visual questions according to our proposed model. In comparison, when using the ReLU activation function [37] and also utilizing 2 glances, the accuracy achieved was 78.14%. Similarly, with the ReLU activation function and 4 glances, the accuracy reached 78.46%. When employing the tanh activation function and 4 glances, the accuracy obtained was 77.99%. Based on the results, it can be concluded that the tanh activation function outperforms ReLU in terms of optimal loss rate. Despite the slightly weaker performance with 4 glances, the tanh function with 2 glances yields the most optimal outcome for answering visual questions in our proposed model.

4- 4- The Results of the Branch Based on Foreign Knowledge

To assess the performance of retrieving external knowledge, we use a metric called R@k, which represents the percentage of question-answer pairs where the original comment is ranked within the top k positions. The basic model employs a two-stage approach for external knowledge retrieval, which yields the highest performance. The initial step utilizes a comprehensive approach involving TF-IDF, preprocessing (PP), and n-grams, resulting in over 90% of question-answer pairs having their original comments ranked within the top 10 positions. The second stage involves re-ranking the comments using a BERT-based model.

Within the knowledge-based branch, specifically in the experimental segment focused on knowledge-based questions, we successfully found the correct answers for 1,730 out of 3,626 questions. This equates to a 47.7% accuracy rate for obtaining the correct answers. For a detailed breakdown of the outcomes in this branch, please refer to Table 3.

4- 5- Final Result

In the AQUA test division, we conducted an extensive examination comprising a total of 4912 questions. Notably, we made significant advancements in the visual branch by substituting the MUTAN attention module with the MLB attention module. This modification enabled us to effectively address 1015 visual questions, resulting in a remarkable accuracy improvement from 77.76% to 78.92% when compared to the visual branch of the base model.

Moreover, our efforts extended to the knowledge-based branch, where we successfully answered 1730 questions. Combining the outcomes from both branches, our proposed model achieved an overall accuracy of 55.88%. These remarkable findings are presented in Table 4, showcasing the final results obtained through our model's implementation.

4- 6- Comparison of the proposed method with other methods

Table 5 presents a comprehensive performance comparison of our proposed model with various other models on the experimental division of AQUA. It is worth noting that our model, referred to as "Ours," outperforms not only the basic model called "VIKING" but also other renowned models. This superiority can be attributed to the significant enhancements we have incorporated into the visual branch of our proposed model.

Within Table 5, different abbreviations are utilized to indicate the input modalities used by each model. Specifically, "P" denotes the utilization of pictures, "K" represents the usage of knowledge, "Q" signifies the involvement of questions, and "w/o" indicates the absence of a specific input modality. Notably, the LSTM [36], BERT, and XLNet methods solely rely on queries to respond to dataset queries. On the other hand, the BAN model [37] leverages both drawings and questions to answer inquiries effectively.

By conducting this comparative analysis in Table 5, we not only demonstrate the superior performance of our

Table 4. The final results obtained by the proposed model

Branch	Correct Answers	Branch questions	Questions	EM Accuracy
visual	1015	1270	4912	0.2066
knowledge	1730	3626	4912	0.3521
total result	2745	4912	4912	0.5581

Table 5. Comparison of the accuracy obtained in different models

Method	Q	P	K	EM
LSTM	✓	-	-	0.198
BERT	✓	-	-	0.194
XLNet	✓	-	-	0.193
BAN	✓	✓	-	0.224
VIKING w/o K	✓	✓	-	0.204
VIKING w/o P	✓	-	✓	0.352
VIKING full	✓	✓	✓	0.555
Ours w/o K	✓	✓	_	0.206
Ours w/o P	✓	_	✓	0.3521
Ours full	✓	✓	✓	0.5588

proposed model but also shed light on the specific modalities and approaches adopted by other models in the field.

5- Future Studies

In the “Future Studies” section of this paper, we turn our attention toward potential avenues for further research and exploration in the domain of visual question-answering. As we delve into the uncharted territories of future investigations, several intriguing questions emerge that not only expand our understanding but also pave the way for innovative developments.

One question that might be intriguing is what happens if the number of questions increases from 10 to 20? Increasing the number of questions from the current top 10 to 20 in our question-answering system, which utilizes a pre-trained BERT model for answer prediction and a fine-tuned XLNet for external knowledge retrieval from the AQUA dataset, could have a significant impact on both the results and performance of the model. Currently, our model achieves a 47.7% accuracy rate for obtaining correct answers to knowledge-based questions. By increasing the number of questions, the model would be required to retrieve and rank more external knowledge, which might increase the complexity of the task. This could potentially lead to a decrease in the accuracy rate if the model struggles to effectively prioritize and retrieve relevant information for a larger set of questions. Additionally, the metric R@k used to evaluate the performance may also be affected, as the model’s ability to rank the original comments within the top k positions could be challenged by the increased volume of questions, potentially impacting the overall quality of the results. Therefore, it is essential to carefully consider the trade-offs between increased question volume and maintaining or improving the model’s accuracy and performance in knowledge retrieval and answer prediction in future works.

Also, it seems that the convergence of the results for more than 5 questions occurs around 50 epochs. What are the possible factors contributing to this convergence? The convergence of results occurring around 50 epochs, as observed in the validation accuracy graph for visual question-answering, can be attributed to several factors. One key factor is the choice of activation function, where the use of the tanh activation function during training, combined with 2 glances at the image during the attention mechanism, proved to be optimal for achieving higher accuracy in answering visual questions. This suggests that the tanh function facilitates more efficient learning in the model compared to ReLU. Additionally, the limited improvement in accuracy with 4 glances indicates a diminishing return on increasing attention, further supporting the 2-glance configuration as the most effective. The observation of a flat loss curve after 50 epochs may suggest that the model has largely learned the relevant patterns and information from the training data, leading to a stable performance plateau. However, a more thorough and comprehensive investigation in this area, especially across different datasets, is necessary to pinpoint the specific reasons behind this phenomenon.

6- Conclusion

In conclusion, this article presented a model for answering visual questions in the context of artistic images by incorporating external knowledge. The study focused on enhancing the performance of a visual question-answering system using a specialized dataset, which contributes to the practicality and applicability of such systems in different fields. By improving upon the VIKING model, the proposed model achieved better accuracy compared to other prominent models and the basic model. Specifically, the accuracy in the visual branch increased from 77.76% to 78.92%, while the overall accuracy across both branches improved from 55.5% to 55.88%.

Future work can involve applying changes to the knowledge-based branch to further increase its accuracy by utilizing more optimal pre-training models. Additionally, incorporating alternative iQAN models in the visual branch, enlarging the dataset, and diversifying the range of questions can contribute to further improvements in the field. These enhancements will enable the system to effectively answer a higher number of knowledge-based questions in the AQUA dataset.

By addressing the limitations of previous research that focused on generic datasets, this study demonstrates the importance of considering specialized domains like art. The combination of computer vision and natural language processing techniques opens up possibilities for practical applications in areas such as medicine and art, where visual understanding and knowledge-dependent reasoning are crucial. By utilizing the AQUA dataset, which incorporates both visual content and associated opinions, this study provides a valuable contribution to the field of visual question-answering in the context of art images.

The proposed model, with the integration of the MLB attention mechanism and the MUTAN fusion mechanism, showcases the potential for improving the performance of visual question-answering systems. Further investigations into the impact of parameters, such as the number of glances and activation functions, can lead to additional enhancements in the model's accuracy and efficiency.

In summary, this research contributes to the advancement of visual question-answering systems by addressing the challenges specific to the domain of art images. By incorporating external knowledge and leveraging specialized datasets, the proposed model demonstrates improved accuracy and lays the foundation for future developments in the field of visual question-answering in specialized domains.

References

[1] Falomir, Zoe, et al. “Categorizing paintings in art styles based on qualitative color descriptors, quantitative global features and machine learning (QArt-Learn).” Expert Systems with Applications 97 (2018): 83-94.

[2] Deng, Yingying, et al. “Exploring the representativity of art paintings.” IEEE Transactions on Multimedia 23 (2020): 2794-2805.

[3] Ma, Daiqian, et al. “From part to whole: who is behind the painting?.” Proceedings of the 25th ACM international conference on Multimedia. 2017.

[4] Rodriguez, Catherine Sandoval, Margaret Lech, and Elena Pirogova. “Classification of style in fine-art paintings using transfer learning and weighted image patches.” 2018 12th International Conference on Signal Processing and Communication Systems (ICSPCS). IEEE, 2018.

[5] Huang, Ru. “Research on Classification and Retrieval of Digital Art Graphics Based on Hollow Convolution Neural Network.” 2022 International Conference on Artificial Intelligence and Autonomous Robot Systems (AIARS). IEEE, 2022.

[6] N. García, B. Renoust, and Y. Nakashima, “Context-aware embeddings for automatic art analysis,” 2019.

[7] N. Huckle, N. García, and Y. Nakashima, “Demographic influences on contemporary art with unsupervised style embeddings,” ArXiv, vol. abs/2009.14545, 2020.

[8] Matsuo, Shin, and Keiji Yanai. “CNN-based style vector for style image retrieval.” Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval. 2016.

[9] Liu, Dilin, and Hongxun Yao. “Artistic image synthesis with tag-guided correlation matching.” Multimedia Tools and Applications (2023): 1-12.

[10] Fumanal-Idocin, Javier, et al. “ARTxAI: Explainable Artificial Intelligence Curates Deep Representation Learning for Artistic Images using Fuzzy Techniques.” arXiv preprint arXiv:2308.15284 (2023).

[11] N. García and G. Vogiatzis, "How to read paintings: Semantic art understanding with multi-modal retrieval," 2018.

[12] N. Garcia et al., "A dataset and baselines for visual question answering on art," CoRR, vol. abs/2008.12520, 2020.

[13] Y. Li, N. Duan, B. Zhou, X. Chu, W. Ouyang, and X. Wang, "Visual question generation as dual task of visual question answering," CoRR, vol. abs/1709.07192, 2017.

[14] Z. Yang, Z. Dai, Y. Yang, J. G. Carbonell, Ruslan Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," CoRR, vol. abs/1906.08237, 2019.

[15] J.-H. Kim, Kyoung Woon On, W. Lim, J. Kim, J. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling," CoRR, vol. abs/1610.04325, 2016.

[16] H. Ben-Younes, R. Cadène, M. Cord, and N. Thome, "MUTAN: Multimodal tucker fusion for visual question answering," CoRR, vol. abs/1705.06676, 2017.

[17] Chen, Kan, Wang, Jiang, Chen, Liang-Chieh, Gao, Haoyuan, Xu, Wei, and Nevatia, Ramakant. Abc-cnn: An attention based convolutional neural network for visual question answering. ArXiv, abs/1511.05960, 2015.

[18] Wang, Peng, Wu, Qi, Shen, Chunhua, Dick, Anthony R., and van den Hengel, Anton. Fvqa: Fact-based visual question answering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40:2413–2427, 2018.

[19] Shih, Kevin J., Singh, Saurabh, and Hoiem, Derek. Where to look: Focus regions for visual question answering. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4613–4621, 2016.

[20] Ilievski, Ilija, Yan, Shuicheng, and Feng, Jiashi. A focused dynamic attention model for visual question answering. ArXiv, abs/1604.01485, 2016.

[21] Xu, Huijuan and Saenko, Kate. Dual attention network for visual question answering. 2017.

[22] Zhu, Chen, Zhao, Yanpeng, Huang, Shuaiyi, Tu, Kewei, and Ma, Yi. Structured attentions for visual question answering. 2017 IEEE International Conference on Computer Vision (ICCV), pages 1300–1309, 2017.

[23] Li, Qing, Tao, Qingyi, Joty, Shafiq R., Cai, Jianfei, and Luo, Jiebo. Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions. ArXiv, abs/1803.07464, 2018.

[24] Wu, Chenfei, Liu, Jinlai, Wang, Xiaojie, and Li, Ruifan. Differential networks for visual question answering. In AAAI, 2019.

[25] Ren, Shaoqing, He, Kaiming, Girshick, Ross B., and Sun, Jian. Faster r-cnn: Towards realtime object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39:1137–1149, 2015.

[26] Patro, Badri N. and Namboodiri, Vinay P. Differential attention for visual question answering. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7680–7688, 2018.

[27] Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." Advances in Neural Information Processing Systems 35 (2022): 23716-23736.

[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," CoRR, vol. abs/1810.04805, 2018.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," CoRR, vol. abs/1512.03385, 2015.

[30] Srimaneekarn, Natchalee, et al. "Binary response analysis using logistic regression in dentistry." International Journal of Dentistry 2022 (2022).

[31] Grootendorst, Maarten. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure." arXiv preprint arXiv:2203.05794 (2022).

[32] Z. Dai, Z. Yang, Y. Yang, J. G. Carbonell, Q. V. Le, and Ruslan Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," CoRR, vol. abs/1901.02860, 2019.

[33] Y. Goyal, Tejas 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," CoRR, vol. abs/1612.00837, 2016.

[34] A. Singh et al., "MMF: A multimodal framework for vision and language research," 2020.

[35] M. Heilman and N. A. Smith, "Good question! Statistical ranking for question generation," pp. 609–617, Jun. 2010.

[36] X. Du, J. Shao, and C. Cardie, "Learning to ask: Neural question generation for reading comprehension," CoRR, vol. abs/1705.00106, 2017.

[37] W Malfliet, "The tanh method: a tool for solving certain classes of nonlinear evolution and wave equations," Journal of Computational and Applied Mathematics, vol. 164–165, pp. 529–541, 2004.

[38] Abien Fred Agarap, "Deep learning using rectified linear units (ReLU)," CoRR, vol. abs/1803.08375, 2018.

[39] Ghojogh, Benyamin, and Ali Ghodsi. "Recurrent Neural Networks and Long Short-Term Memory Networks: Tutorial and Survey." arXiv preprint arXiv:2304.11461 (2023).

[40] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear attention networks," CoRR, vol. abs/1805.07932, 2018.

HOW TO CITE THIS ARTICLE

E. Zolghadriha, K. Fouladi-Ghaleh, P. Ardehkhani. Envisioning Answers: Unleashing Deep Learning for Visual Question Answering in Artistic Images. AUT J Electr Eng, 56(2) (2024) 191-202.

DOI: 10.22060/eej.2023.22605.5552