Art Visual Question Answering
(Deep Learning Term Project)

Ans Munir
MS Data Science
ITU, Lahore
msds20033@itu.edu.pk

Abstract

It's challenging to respond to questions regarding artistic paintings. It denotes comprehension of the visual data portrayed in the artwork as well as logical information gathered via research into the historical context of craft. In this paper, we provide AQUA (Art Question Answering), our first attempt at creating a new dataset. Question-Answer (QA) pairings are constructed automatically using cutting-edge question generation methods based on paintings and feedback offered in a previously collected set of art comprehension data. Crowd-sourced workers clean up the QA pairs for grammatical correctness, responsibility, and answer accuracy. Our dataset includes visual questions based on the painting as well as knowledge based on feedback. As a foundation, we also present a two-pronged strategy in which visual and knowledge challenges are treated independently. We compare our standard model to cutting-edge models for question answering, and we provide a detailed report on the problems and potential future bearings for visual inquiry replying on art.

Original Paper Reference

I presented the research paper in the class as a term project for my deep learning course, reproduced the results by running its code and have written a report on it. I am not the author or co-author of this paper. The original paper and the authors can be found on this Link[1]

1. Introduction

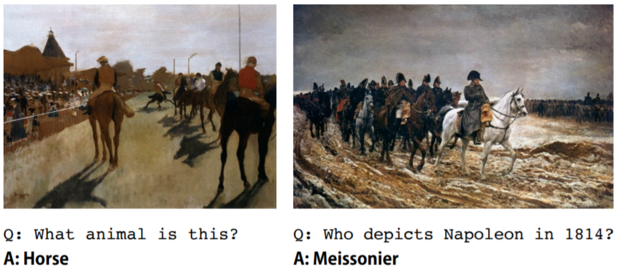


Figure 1. Examples from the AQUA dataset.There are two different types of QA pairs:generated from paintings (left) and generated from paintings' comments (right). Source: Original Paper

Visual Question Answering (VQA) is a popular topic [29] and one of the problems of semantic understanding of visual input. VQA necessitates a more in-depth analysis of visual content, as well as inquiries provided by humans in natural language. VQA task extensions now include knowledge-based tasks, [30]. The primary goal of VQA tasks that depict natural world items, locations, and events is the natural image. Very few studies discuss the many types of information that are visually expressed, for example, the VQA dataset of abstract image subset [1], CLEVR [17], and PororoQA [18]. The primary objective for employing non-real-world images is to free up visual recognition in the VQA pipeline, allowing it to focus on response prediction. Paintings, on the other hand, are a fascinating domain for VQA. Aside from their conventional and historical value, paintings offer difficulties in VQA:

- First, the subject of the painting can be depicted in various abstractions, such as naturalism, realism, symbolism, impressionism, cubism, and so on. As a result, pre-trained models for object identification, for example, may be employed for realism but are not a good answer for cubism.
- Second, background knowledge and its interpretations are important, such as the historical and personal context of the author, which is not conveyed in the picture. This shows that historical knowledge of the history of painting may be required to address the problem and successfully answer inquiries.

In this research, we examined the outcomes of the AQUA

[1] https://arxiv.org/pdf/2008.12520.pdf

(art question answering) dataset, which is based on the intelligent dataset [12], as well as baseline models. The QA pairs for this study (shown in Figure 1) are generated automatically by combining different question generating approaches from natural language processing [6] and computer vision [17]. Given the Se-mArt dataset, we used both the paintings and the comments as input to generate QA pairs. Thus, we construct visual questions (e.g., ”What animal is this?” in Figure 1) from the information provided in these paintings, and knowledge-based questions (e.g., ”Who depicts Napoleon in 1814?” in Figure 1) using comments. To address this new QA problem, we employed VIKING (visual- and knowledge-branch network for predicting answers), a baseline model created particularly for art knowledge information that employs modality selection on top of previous QA and text QA models. With specific branches, this network addresses dual-modality, i.e., visual and external knowledge-based inquiries.

Initially, we looked at art question responding, which needs visual perception of paintings and background knowledge. Textual content, such as Wikipedia, books, and so on, can provide background knowledge. This domain of answering questions using visual and contextual information has clearly not been explored.

Second, we conducted our study using the AQUA dataset. The QA pairs in this dataset were manually purified by crowdsourcing workers in terms of answerability, grammatical correctness, and answer correctness.

Finally, for our art QA work, we used a baseline model called VIKING. The baseline models, in addition to the question, employ the artwork and a paragraph collected from a knowledge base to predict the answer applicable to both the question and the painting.

2. Literature Review

Considering every artwork contains some visual information, arts and computer vision are inextricably linked. A fundamental focus is on digitising works of art for archival and therapeutic purposes. Several studies in the field of computer vision have been conducted for works of art, including authorship/style identification [28], image classification [10], and image retrieval [3]. This is the very first attempt in the history of VQA to answer questions about paintings.

2.1. Question Generation

The dataset for Question Answering on Real-world Images (DAQUAR) was released by Malinowski and Fritz [22]. They developed a system that blends semantic parsing with picture segmentation. Their approach is notable as one of the first attempts at image QA, although it has numerous shortcomings. As a result, it is an expensive operation with larger datasets. [23]

There are two types of visual question generation (VQG): grounded and open-ended [26]. Grounded VQG asks and answers questions about the input image [31]. So captions are constructed using the image, and we can generate questions from the captions using either rule-based [27] or neural [31] models. Open-ended VQG is concerned with abstract ideas such as occurring and situations deduced from visual objects [25]. The importance of diversity in open-ended VQG cannot be overstated. Incorporating variational auto-encoders [15] and generative adversarial networks [23] helps with this.

A rule-based or neural model-based approach has been used for Textual Question Generation (TQG). Initially, rule-based TQG builds question templates manually [14] or by crowd-sourcing [20] and then generates questions by applying these templates. [6] The first pioneering neural model for TQG, which uses the sequence-to-sequence model, was presented. The study of neural TQG focuses on how to encode replies [19], produce question words [7], and employ paragraph-level context [5].

2.2. Visual Question Answering

Previous VQA research has used either natural photos or videos. Image-based VQA approaches include joint visual and language embedding, as well as attention mechanisms that illustrate how we might incorporate spatial attention into the conventional LSTM model [29]. Because it contains time information, video-based question answering is an extension of VQA. We can also include action recognition [16], story comprehension [18], and temporal coherence [32]. Another intriguing expansion of VQA is external knowledge beyond images and videos. The information could be generic [24] or dataset-specific [9].

Because dataset collection is not the major focus of VQA, we mostly employ synthetic datasets. Malinowski and Fritz [22] built their dataset autonomously using various templates, while Johnson et al. [17] likewise used the synthetic dataset to reduce probable biases from the conventional dataset. We are also utilising a synthetic dataset to demonstrate the concept of the VQA in the realm of art.

3. AQUA Dataset

We used an existing dataset, SemArt dataset [11], to generate the AQUA dataset. This dataset includes the painting as well as any comments or metadata associated to the artwork, such as a block of text or the author's name. As a result, we will use the visual image and the knowledge supplied in the comments for the QA pair in our dataset. As a result, the question creation module has two modules that generate questions based on the knowledge provided in words and the visual element. For optical question creation, we will employ the cutting-edge model iQan [21]. IQan required visual data, which will be our painting, as well as an answer that can be predicted by detecting items in an image using Amazon Rekognition.

Aside from that, Pythia can yield the same results when producing captions from images and using captions generated from images. We can establish a QA pair using a rule-based approach, such as TQG, which generates a parsing tree from the provided text and generates a QA pair. TQG will be used again for question generation from knowledge, but this time it can be neural or rule-based. So, once we have generated the QA pairings from both sources, we will evaluate them.

For their grammatical correctness, whether they have answers for the specific question, whether the provided solution is correct for the specific question, what we need to answer the specific question, whether textual or visual information, and, finally, whether the question is reasonable enough to ask. We can define any problem on the dataset after it has been cleansed, but for the time being, we are focusing on addressing the question from either the visual or knowledge branch.

4. Methodology

There are two sorts of questions in the AQUA dataset: visual questions and knowledge-based questions. We created the VIKING model, which has two specialised branches to handle each sort of question individually. Our model is made up of three parts:

- Modality Selector
- Visual QA Branch
- Knowledge QA Branch

4.1. Modality Selector

Modality Selector S determines if the question is visual or knowledge-based and routes it to the appropriate branch. The question is encoded first by a BERT, which generates a 1024-dimensional vector q. Along with the question, we included the painting. The painting will be encoded into a 2048-dimensional vector v using ResNet-152. The vectors v and q will be concatenated into vector x, and then a logistic regression model will be trained.

$$S(x) = \frac{1}{1 + e^{-(w_s^T x - b_s)}} \tag{1}$$

where w_s and b_s are a scalar and trainable vector . When S(x) > 0.5 question q is passed to the visual QA branch and to the knowledge QA branch otherwise.

4.2. Visual Question Answer Branch

Without any outside knowledge, visual questions can only be answered based on the paintings themselves. In this case, iQAN was employed in the visible question branch to answer the given question. iQAN is a module that may be used to answer visual questions as well as generate visual questions. To achieve the best results, iqan employs MUTAN [2], a multi-modal fusion technique based on Tucker decomposition. We trained iQAN on AQUA data set train data and did not use a previously trained model for question generating. The Visual QA branch returns an answer a_v based on the vocabulary A's 5000 most common terms in training data.

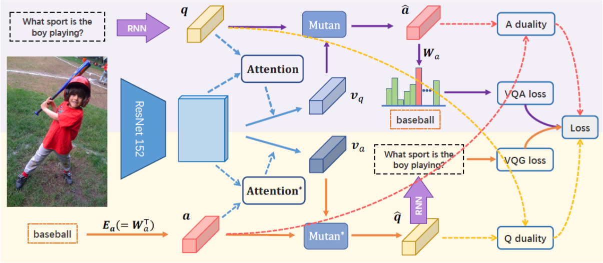


Figure 2. Invertible Question Answering Network (iQAN) consists of VQA and VQG respectively. Source: Original Paper

4.3. Knowledge Question Answer Branch

Any questions classified as Knowledge-based by the modality selector are routed to the Knowledge QA branch. We apply a two-stage approach in the knowledge QA branch to identify the most relevant comment from all the comments C. First, we rank all of the comments for a given question using TF-IDF. Then we select a subset C_q that includes the top ten most relevant comments. We re-ranked the comments in subset C_q using a pre-trained BERTcitebert model in the second step to discover the most relevant comment c_q. This is converted into a sentence-pair classification strategy by me, which predicts the relevance of the comment and the query. This two-stage strategy will significantly reduce computing costs. The output of the BERT model is then fed into the XL-Net model [8]. The starting and finishing points of the answer are then determined by XL-Net based on the query. This process can be seen from the Figure 2 [13] and Figure 4

4.4. Teaser Diagram

The teaser diagram for our model has been shown in the figure 4. The image and the question will be taken by our model. The modality selector will determine whether the question is visual or knowledge-based and will route it to the appropriate component. In the Visual component, we used IQAN to anticipate the answer from the image.

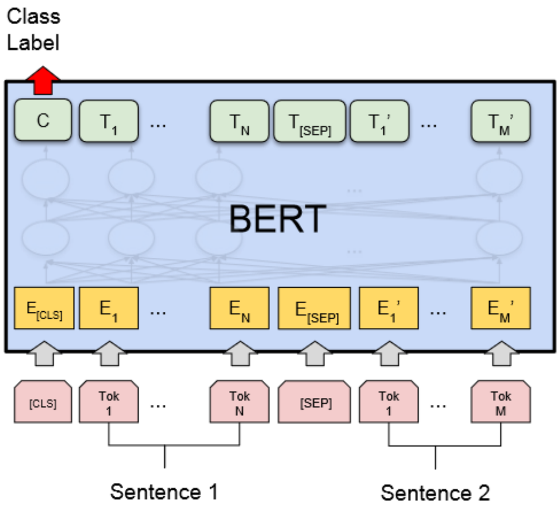


Figure 3. BERT model for sentence pair similarity task [4]

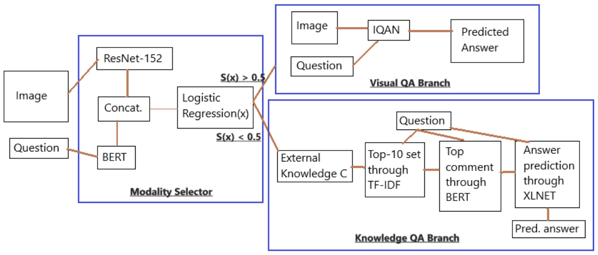


Figure 4. Teaser Diagram showing the overall working of the model.

5. Results & Discussion

The VIKING models include three fundamental components, which the current study reviewed and analysed in order to gain a better understanding of the domain and the possibilities of multiple areas of breakthrough.

The first stage, Modality selector, predicts if external knowledge is required to answer a question given its corresponding artwork. This model employs Pre-trained ResNet-152 for painting feature preparation and BERT for question feature preparation. We developed a logistic regression model with 99.6% accuracy.

The second stage, also known as the Visual QA branch, predicts a response for a sample categorised as not requiring external information. For image feature preparation, we utilised resnet152_size,448. We trained the iQAN model to answer visual questions. The Figures 5 and 6 demonstrate the loss and accuracy curves for visual branch training, respectively.

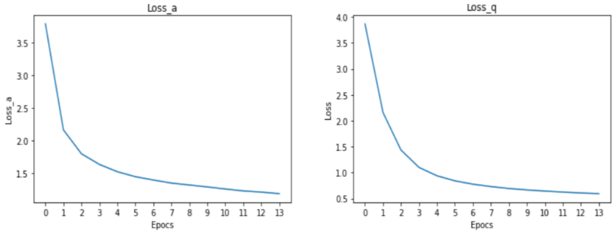


Figure 5. The diagram shows training loss curve for visual branch (Left: Loss curve for answer generation branch, Right: Loss curve for question generation branch)

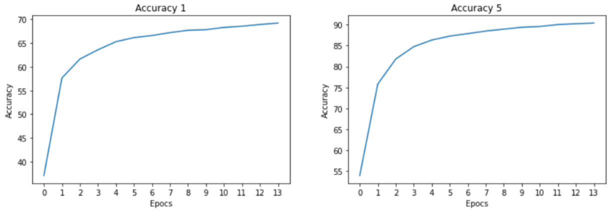
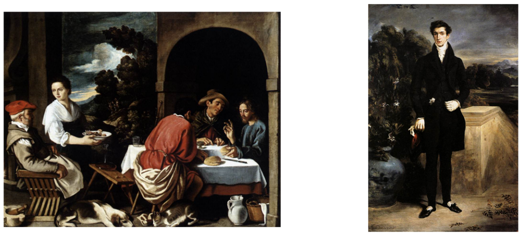


Figure 6. The following diagram shows the training accuracy curve for the visual branch (Left: Accuracy curve based on top 1 prediction, Right: Accuracy curve based on top 5 predictions)



```
question: "what is the man sitting on"
answer: "chair"
need_external_knowledge:     false
external_knowledge_predict:  false
```

```
question: "what had delacroix sold to th
e french state"
answer: "his picture of the Massacre of
Chios"
need_external_knowledge:     true
external_knowledge_predict:  true
```

Figure 7. The figure shows the prediction of modality selector to predict if the external information is required to answer the input question

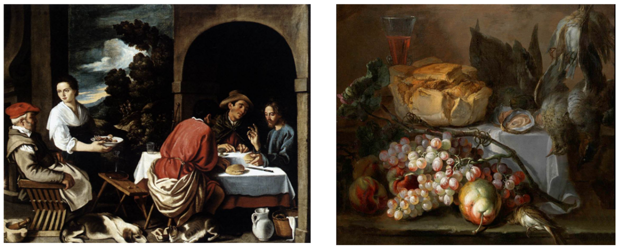
The output of two photos where the modality selector accurately anticipated whether external information is necessary to answer the input question is shown in the Figure 7. Questions that do not require external information are shifted to the knowledge branch rather than the visual branch.

The output of the question where the modality selector classified the question as visual information not required is shown in the Figure 8.

The Figure 9 depicts the knowledge branch predictions for the images where the modality selector anticipated that the knowledge base information is required to solve the query. It can be seen from the offered answers that these answers are not very precise due to the unknown visual information for the knowledge branch.

6. Conclusion

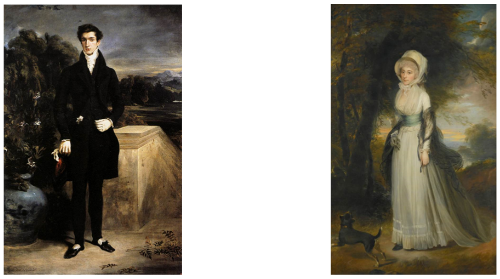
This research develops a new dataset based on an intelligent dataset that has been available in the VQA business for some time. In addition, we offered a new baseline for



```
question: "what is the man sitting on"
answer: "chair”
predict_answers: ["chair", "couch"]
```

```
question: "wwhat are all on a stone"
answer: “oyster”
predict_answers: ["people", "man"]
```

Figure 8. The visual branch output is shown in the figure



```
question: "what had delacroix sold to the
french state"
answer: "his picture of the Massacre of
Chios”
predicted_answer: "he"
```

```
question: "where does the painting stand"
answer: "in a landscape"
predicted_answer: "painting within a
painting"
```

Figure 9. The knowledge branch output is shown in the figure. It can be observed from the output that the without the visual context these answers are not accurate.

responding questions about the art, which includes three modules, one of which is a modality selector, which picks the module through which we can acquire the answer to the question. The chosen module is then used to answer the query. Because the answer is now developed utilising one of the branches, either visual or knowledge, we can attain better results by combining both concurrently. Once the answer has been influenced by both branches, we can concatenate them for improved output.

7. Future Improvements

Our model is currently separated into two components. The first is a visual component, while the second is a knowledge component. We have a modality selector that guides images to either the visual or knowledge branches. However, we wish to answer the question using both branches. For this, we are producing questions that require both the visual and knowledge parts of the response to fail the current model, and then we will try to make modifications in this model that will require both the visual and knowledge parts of the answer to fail the current model.

References

[1] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. Vqa: Visual question answering, 2016. 1

[2] Hedi Ben-younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering, 2017. 3

[3] E.J. Crowley and Andrew Zisserman. The state of the art: Object retrieval in paintings using discriminative regions. 01 2014. 2

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 4

[5] Xinya Du and Claire Cardie. Harvesting paragraph-level question-answer pairs from Wikipedia. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1907–1917, Melbourne, Australia, July 2018. Association for Computational Linguistics. 2

[6] Xinya Du, Junru Shao, and Claire Cardie. Learning to ask: Neural question generation for reading comprehension, 2017. 2

[7] Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. Question generation for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 866–874, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. 2

[8] Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*, 2019. 3

[9] Noa Garcia, Mayu Otani, Chenhui Chu, and Yuta Nakashima. Knowit vqa: Answering knowledge-based questions about videos. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:10826–10834, Apr. 2020. 2

[10] Noa Garcia, Benjamin Renoust, and Yuta Nakashima. Context-aware embeddings for automatic art analysis. *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, Jun 2019. 2

[11] Noa Garcia and George Vogiatzis. How to read paintings: Semantic art understanding with multi-modal retrieval. 2018. 2

[12] Noa Garcia and George Vogiatzis. How to read paintings: Semantic art understanding with multi-modal retrieval. In Stefan Roth and Laura Leal-Taixé, editors, *Computer Vision – ECCV 2018 Workshops, Proceedings*, volume 11130 of *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pages 676–691, Germany, Jan. 2019. Springer. 15th European Conference on Computer Vision, ECCV 2018 ; Conference date: 08-09-2018 Through 14-09-2018. 2

[13] Noa Garcia, Chentao Ye, Zihua Liu, Qingtao Hu, Mayu Otani, Chenhui Chu, Yuta Nakashima, and Teruko Mitamura. A dataset and baselines for visual question answering on art. In *Proceedings of the European Conference in Computer Vision Workshops*, 2020. 3

[14] Michael Heilman and Noah A. Smith. Good question! statistical ranking for question generation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Lin-*

guistics, pages 609–617, Los Angeles, California, June 2010. Association for Computational Linguistics. 2

[15] Unnat Jain, Ziyu Zhang, and Alexander Schwing. Creativity: Generating diverse questions using variational autoencoders, 2017. 2

[16] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering, 2017. 2

[17] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning, 2016. 1, 2

[18] Heo M.O. Choi S.H. Zhang B.T.: Kim, K.M. : Video story qa by deep embedded memory networks, 2017. 1, 2

[19] Yanghoon Kim, Hwanhee Lee, Joongbo Shin, and Kyomin Jung. Improving neural question generation using answer separation, 09 2018. 2

[20] Igor Labutov, Sumit Basu, and Lucy Vanderwende. Deep questions without deep understanding. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 889–898, Beijing, China, July 2015. Association for Computational Linguistics. 2

[21] Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, and Xiaogang Wang. Visual question generation as dual task of visual question answering. 2017. 2

[22] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input, 2015. 2

[23] Mateusz Malinowski and Mario Fritz. Towards a visual turing challenge, 2015. 2

[24] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge, 2019. 2

[25] Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. Generating natural questions about an image, 2016. 2

[26] Liangming Pan, Wenqiang Lei, Tat-Seng Chua, and Min-Yen Kan. Recent advances in neural question generation, 2019. 2

[27] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering, 2015. 2

[28] Lior Shamir. Computer analysis reveals similarities between the artistic styles of van gogh and pollock. *Leonardo*, 45:149–154, 04 2012. 2

[29] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Visual question answering: A survey of methods and datasets, 2016. 1, 2

[30] Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Ask me anything: Free-form visual question answering based on knowledge from external sources, 2016. 1

[31] Shijie Zhang, Lizhen Qu, Shaodi You, Zhenglu Yang, and Jiawan Zhang. Automatic generation of grounded visual questions, 2017. 2

[32] Linchao Zhu, Zhongwen Xu, Yi Yang, and Alexander G. Hauptmann. Uncovering temporal context for video question and answering, 2015. 2