The current and future role of visual question answering in eXplainable artificial intelligence.

CARO-MARTINEZ, M., WIJEKOON, A., DIAZ-AGUDO, B. and RECIO-GARCIA, J.A.

2023

The Current and Future Role of Visual Question Answering in eXplainable Artificial Intelligence

Marta Caro-Martínez[1,*], Anjana Wijekoon[2], Belén Díaz-Agudo[1] and Juan A. Recio-García[1]

[1] *Department of Software Engineering and Artificial Intelligence, Universidad Complutense de Madrid, Spain*
[2] *School of Computing, Robert Gordon University, Aberdeen, Scotland*

Abstract
Over the last few years, we have seen how the interest of the computer science research community on eXplainable Artificial Intelligence has grown in leaps and bounds. The reason behind this rise is the use of Artificial Intelligence in many daily life tasks, and the consequent necessity of people to understand the intelligent systems' behaviour. Computer vision-related tasks are not an exception, for example, Visual Question Answering tasks. The Artificial Intelligence models that carry out this specific task make an effort to answer questions about what we can watch in a particular image. In this work, we review the existing work about eXplainable Artificial Intelligence on Visual Question Answering which is a problem on which there is still much work to be done. Moreover, we open the discussion about the challenges to overcome regarding this topic, like the future role of Visual Question Answering to address eXplainable Artificial Intelligence issues or difficulties.

Keywords
eXplainable Artificial Intelligence, XAI, Visual Question Answering, VQA, Natural Language Processing, NLP, Computer Vision

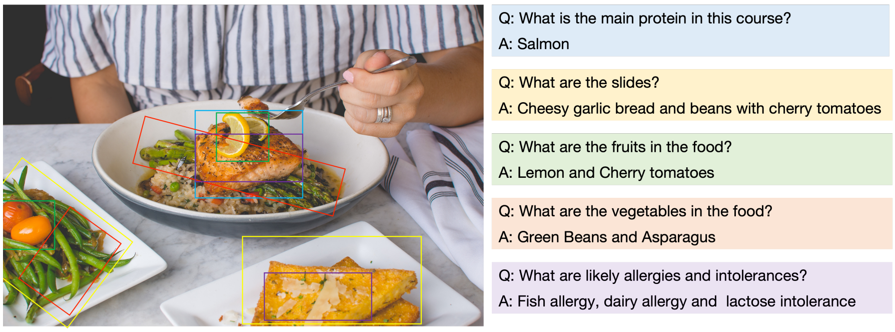
1. Introduction

Artificial Intelligence (AI) has become an integral part of human life due to the high volume of research targeted towards helping people perform many complex tasks. For example, there is a huge interest in making machines observe, understand and perform multiple tasks with images, i.e. machine vision. Machine vision-related tasks have multiple useful applications in many domains, like detecting breast cancer, detecting manufacturing industrial defects, or helping disabled people [1]. Furthermore, AI research has extended to reasoning with multiple modalities including images. One of these and the focus of this paper is Visual Question Answering (VQA), which involves reasoning with computer vision and natural language [2]. AI models for VQA try to answer questions about the content of a specific image (an example is shown in Figure 1). Moreover, VQA shares the common challenge with other AI tasks that it lacks explainability, which affects the users' trust and the system's performance as a consequence.

ICCBR XCBR'23: Workshop on Case-based Reasoning for the Explanation of Intelligent Systems at ICCBR-2023, July, 2023, Aberdeen, Scotland

[*] Corresponding author.

martcaro@ucm.es (M. Caro-Martínez); a.wijekoon1@rgu.ac.uk (A. Wijekoon); belend@ucm.es (B. Díaz-Agudo); jareciog@ucm.es (J. A. Recio-García)



Q: What is the main protein in this course?
A: Salmon

Q: What are the slides?
A: Cheesy garlic bread and beans with cherry tomatoes

Q: What are the fruits in the food?
A: Lemon and Cherry tomatoes

Q: What are the vegetables in the food?
A: Green Beans and Asparagus

Q: What are likely allergies and intolerances?
A: Fish allergy, dairy allergy and lactose intolerance

Figure 1: Examples of VQA: each question is answered by identifying relevant information in the image. Each coloured box indicates the regions considered in answering the question (best viewed digitally in colour).

In recent years, eXplainable Artificial Intelligence (XAI) is one of the most popular research fields on AI [3, 4]. This popularity has risen due to the increment of AI systems applications in daily life, and especially in critical domains like healthcare, security, or industry. Another reason is the increasing complexity of AI algorithms and methodologies. For instance, previous approaches such as rule-based systems and decision trees [5] were simple and importantly interpretable [3]. However, the increasing complexity of more recent gradient-based models and algorithms, commonly referred to as black boxes, renders them non-interpretable for end-users [3, 6].

In the literature, VQA has been achieved using both knowledge-light [7, 8] and knowledge-intensive methods [9]. In this paper, we explore the explainability of both approaches due to the necessity of including XAI in VQA to increase users' trust on these approaches. While knowledge-intensive approaches may seem interpretable, it is worth exploring due to the increasing demand for explainability and the lack of existing review in this research area. Accordingly, this paper will study and review the existing work on applying XAI techniques to make VQA more explainable. Additionally, we want to depict challenges and future work lines that are going to be necessary to develop in the following years. Specifically, we will explore the future work on XAI applied to explain VQA and also the challenges in XAI research that can be addressed using VQA techniques.

The rest of the paper is organised as follows. Section 2 presents the literature review methodology followed by a review of literature in Visual Question Answering (Section 3). Section 4 presents the existing literature that explored explainability in the context of VQA and we conclude with Section 5 which discusses open challenges and the utility of VQA in XAI. Finally, we end the paper getting some conclusions in Section 6.

2. Literature Review Methodology

To study the literature related to VQA and VQA with explanations we have performed a structured search considering different terms and dates. We used the Google Scholar [1] database.

[1] https://scholar.google.com/

Table 2 shows the keywords regarding the search that we carried out. From the results obtained during the seeking task, we decided to select the most relevant papers to our topic, which included from 1 to 4 papers for each search.

Terms	Dates	# results (aprox.)	Selected	Notes
visual question answering	Any time	1,610,000	2	Only surveys
visual question answering dataset	Any time	155,000	2	Only surveys
visual question answering dataset	After 2019	16,500	4	from top 4
visual question answering xai	Before 2019	2,830	3	from top 3
visual question answering explanations	Before 2019	17,800	2	from top 3
visual question answering explainability	Before 2019	18,100	3	from top 3
visual question answering explanation system	Before 2019	17,800	2	from top 3
visual question answering xai	After 2019	5,690	3	from top 3
visual question answering explanations	After 2019	17,700	4	from top 4
visual question answering explainability	After 2019	17,600	4	from top 10
visual question answering explanation system	After 2019	16,900	1	from top 5
transformer-based vision language modelling for visual question answering	After 2019	17,400	4	State-of-the-art
transformer-based vision language modelling for visual question answering dataset	After 2019	7,350	4	State-of-the-art
transformer-based vision language modelling for visual question answering explainability	After 2019	1,420	3	State-of-the-art
transformer-based vision language modelling for visual question answering explanations	After 2019	7,090	3	State-of-the-art
transformer-based vision language modelling for visual question answering explanation system	After 2019	6,870	3	State-of-the-art
transformer-based vision language modelling for visual question answering XAI	After 2019	262	3	State-of-the-art

Table 1
Literature search carried out to study VQA and VQA with explanations.

The first set of searches were a way to introduce ourselves in the topic and in the VQA algorithms. Once we discovered there were two main types of VQA algorithms (knowledge-light and knowledge-intensive), we started the search for the rest of the terms (related to transformers especially). In the following sections, divided mainly according to knowledge-light and knowledge-intensive classification, we describe and discuss the works selected from the search.

3. Visual Question Answering

Visual Question Answering is a task that emerged from the necessity of answering questions about an image or video. For example, VQA is utilised when helping the visually impaired to understand the content of a photo [1]. VQA as an AI task is complex (compared to other tasks like classification or regression) and requires contributions from two key domains: computer vision and language modelling [2]. It differs from other computer vision tasks in that we do not know the question to answer until run time. In VQA, we can discuss different types of questions regarding the knowledge that we need to access to answer them [10]. According to literature, we have identified the following types of questions in VQA:

- *Object detection.* For example, “Are there dogs in the picture?”.

- *Fine-grained recognition.* For example, "What type of dog breed appears in the picture?".
- *Action or activity detection and recognition.* For example, "Is the dog eating?".
- *Additional knowledge-based reasoning.* For example, "Is the dog breed the favourite dog breed of Queen Elizabeth II?".
- *Commonsense reasoning inference.* For example, "Does the dog love her humans?".

There are two key approaches to implementing VQA: knowledge-intensive and knowledge-light methods. Both approaches typically involve the following steps: 1) object detection in an image with high accuracy with fine-grained details; 2) language comprehension of the question; and 3) compilation of the answer utilising the information from steps 1 and 2, and external knowledge sources (if required) [11]. Step 3 might be the most challenging one due to the complex reasoning we have to carry out. Sometimes we can find that our questions are very easy to answer considering the objects detected from the image. However, often, we need to infer knowledge from other knowledge sources to disambiguate the uncertainties between our question and the objects detected in our image [12]. These uncertainties are caused by open-ended, ambiguous questions that can take multiple answers or ones that require additional information not available on the image [2]. Mainly, the knowledge-intensive methods are the methods that try to overcome this problem (see Section 3.2).

In the following sections, we explore different methodologies used to implement VQA in detail.

3.1. Knowledge-light AI Methods

The state-of-the-art knowledge-light methods for VQA are vision-language (VL) fusion deep neural architectures. They are optimised to learn the representations of individual modalities as well as the alignment between modalities in a fusion architecture.

The alignment is enforced by the multi-modal contrastive loss which is derived from the theories of Mutual Information and Noise-contrastive Estimation [13]. The multi-modal contrastive loss can be calculated in many forms and often uses a combination of the following in the loss function. Implementation of these losses is also enabled by momentum encoding which regulates the weight updates in pre-trained encoders [14].

- *Inter-modal Alignment* forces representations of the matching image, text pairs to be closer in the feature space while pushing apart the representations of unmatched image, text pairs [8].
- *Intra-modal Alignment* forces the encoders to learn semantic differences between matched and unmatched instances of the same modality. For instance, representations of similar images are forced to be closer while representations of different images are pushed apart [7].

The state-of-the-art VQA models are the ALIGN model trained using contrastive loss [15], ALBEF model trained using the inter-modal alignment [8]; VL model by authors of [7] trained using Triple Constrative Loss; VinVL [16] which integrates improved object detection with VL modelling. They all follow similar fusion architectures while differing in training objectives, training data utilised and pre-train/fine-tune tasks.

3.2. Knowledge-intensive AI Methods

In general, we can say that the knowledge-intensive or knowledge-enhanced methods are the ones that need to use external knowledge, from a different knowledge source that it is not the image, and the question, to get an answer [12, 2].

Using additional knowledge sources can be remarkable in VQA tasks, because as we have mentioned previously, not all the questions that we can find are as simple that we can answer them with only that information. For example, we can have an image where several types of food appear. We can also have the following question: "How many types of fruits are in the picture?" (see Figure 1). The algorithm should recognise not only the foods that we have in the image, but identify and understand which foods are fruits and which ones are not. Then, we can also say that knowledge-intensive methodologies overcome one of the biggest problems that deep learning models have since these ones only consider the information that we can find in the training data and in the query [2, 12]. We have to think that training data can scale but never cover all the possible knowledge present in the world. Therefore, it makes sense to use other bigger, various, and deeper knowledge sources to complement our training data. Moreover, deep learning models are limited, they cannot learn all the knowledge. Regarding this point, knowledge-intensive methods help to enhance the knowledge learnt [2]. Then it is reasonable to wonder whether knowledge-intensive methods for VQA are better than deep learning and whether they are also the future of VQA algorithms.

4. Explainability of VQA

It seems in general VQA systems do in fact follow human reasoning: detect objects in the image and establish the relations between those objects [11]. This has been the key to make the reasoning process of these methods understandable to humans. We categorise such approaches as: *interpretable* methods where the VQA is a pipeline of sub-tasks and the outcome of each step is used to explain the answer; *knowledge-light* methods that use architectural changes or optimisation techniques to generate an explanation; and *knowledge-intensive* methods that exploit external knowledge sources and reasoning to generate explanations. In the next sections, we explore these methods.

4.1. Interpretable Methods for VQA

These models are the ones whose execution process is divided into several steps. This way, we can get intermediate results in every step that can be shown to the users in order to make the process understandable. That is why these approaches are considered interpretable, although the methods used in every step are black-boxes, mainly neural networks.

One example approach is the Grounded Visual Question Answering model (GVQA) [17]. They use different algorithms depending on the type of question that we come across (yes/no or non yes/no question). They also divide the model's behaviour process to get the answer in different steps (getting important parts in the image, retrieving concepts from the question, classifying the type of questions or predicting the answer). So, in every step, the model uses different deep learning algorithms. Nevertheless, the model obtains an output in every step, that can be shown to the target user, making more transparent the reasoning process carried out by the model. This transparency makes the model interpretable compared to other models that take the decision in one shot. Another example is the work by Li et al. [18]. In this case, the method divides the process to carry out in three steps: 1) word prediction (a CNN algorithm is used to solve a classification problem, where an image should be associated with a set of words); 2) sentence generation (where a single-layer LSTM algorithm is applied to get the probability of having a specific word in a set of words, taking into account the set of words that we got in step 1; and 3) answer reasoning (the sentence from step 2 and the question are encoded by two LSTM algorithms to get the answer). Authors claim that the system is interpretable because users can watch the results obtained in every step, which are the words and the sentences that describe the image according to the VQA model.

4.2. Knowledge-intensive Methods for VQA with Explainability

Knowledge-intensive methods for VQA use external knowledge sources to provide explanations. The process carried out by these systems is divided into steps, separating the representation from the reasoning. The graph built to represent the knowledge to use to get the answer can be also used as an explanation itself, or used as the knowledge source for an explanation system. Therefore, knowledge-intensive methods are able to provide explainability [19]. Moreover, the advantages that we pointed out about this type of system in Section 3.2 plus their explainability make knowledge-intensive methods a strong candidate in future when implementing explainability for VQA. Next, we describe some examples found in the literature to illustrate these ideas.

The authors of [20] proposed a post-hoc and model-agnostic approach to provide counterfactual explanations for VQA. In particular, they want to check "What is the response of the VQA model if we substitute word X with word Y in question q". The authors delete words or replace words in the questions using a knowledge graph (WordNet[2]) to get the substitute words. They aim to get these words by looking for synonyms, hypernyms, hyponyms, or siblings in the graph between the words. The original question and the modified questions together with their answers (obtained by a specific VQA model) are presented to the user as local explanations. They also present users with global explanations which is VQA model performance comparison between when answering the original questions and the modified explanations.

VLC-BERT [21] is VQA model that uses COMET for explaining its answers. COMET (Commonsense Transformer) is a commonsense reasoning generation transformer model that given a subject and a relation, it predicts a possible object. An example from the authors is if the subject is "taking a nap" and the relation is "causes" a possible object is "have energy". It is trained and tested on ATOMIC [22] and ConceptNet knowledge graphs both of which consist of social commonsense knowledge. Commonsense reasoning extracted from the COMET is used in VLC-BERT to improve answer generation making it knowledge-intensive.

[2] https://wordnet.princeton.edu/

4.3. Knowledge-light Methods for VQA with Explainability

There are also examples of post-hoc explanation approaches, that are independent of the VQA reasoning process and utilise an explanation module that generates visual and/or textual explanations. Most commonly the generative model is a Deep Learning model that annotates the RoIs in the image. The authors of [19] introduced a VQA model that includes an LSTM-based explanation module. It takes the question and the answer as the input and their multinomial distribution on the important concepts to generate an explanation. The explanation is both textual and visual: the image RoIs that explain the answer are marked in colours, and a complementary text explains the reason behind the answer.

More recent transformer-based VQA models generate explanations in an ante-hoc manner. Authors of [23] proposed the CtF framework that extracts keywords from both input image and question to filter information from the RoI of the image and question tokens to produce answers. The semantic reasoning between RoIs and the question keywords is presented as the explanations. A similar approach to learn from different granularity levels for VQA is used by authors of [24]. They use Kullback Leibler divergence as the loss to align between two modules that learn coarse-grained and fine-grained fusion of the image and question. The answers are explained by visualising the attention on RoIs and question words. The authors of [25] extend the CtF framework with an explanation generation module in addition to predicting the answer. A transformer-based decoder model generates the explanation and it is trained along with the answer classification task.

4.4. Evaluating the Explanations for VQA

Evaluation of explanation quality has improved over the years as much as the implementation of the explainability into VQA. This is supported by dedicated datasets created for Explainability in VQA such as VQA-X and VQA-HAT. In addition, some works set out to use a very well-known explainer (for example LIME or GradCAM) to generate explanations from the image and compare those explanations with the one obtained by their own methods [19].

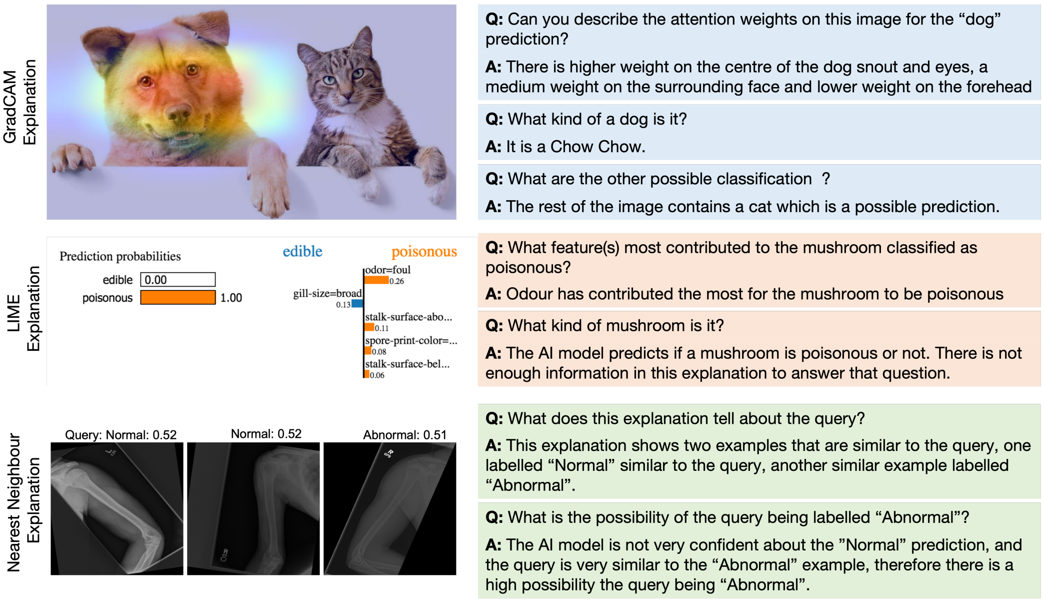
Quantitative evaluation using questionnaires to capture users' opinions on the explanation systems is another common approach we encountered in the reviewed literature. We find the use of questionnaires with Likert scales [26] that is similar to what we encounter in other XAI evaluation domains. They can be used to allow users to pick what is preferred explanation approach (between proposed and a baseline from the literature). Following is a list of existing evaluation metrics and novel metrics proposed to evaluate the explanations for VQA:

BLEU-4 [27]. It is a well-known standard sentence-comparison metric. In this case to evaluate textual explanations [19, 25]. Originally, it is evaluated to check the performance in machine translation. *METEOR* [28]. It is a metric that aims the same goal than BLEU-4 [19]. *ROUGE-L* [29]. This metric tries to measure the quality of summaries created by machines. However, it is being used in the evaluation of explanations for VQA [19, 25]. *CIDEr* [30]. CIDEr is an actual metric to measure VQA [19]. It provides a score about the quality of a sentence that describe a specific image. *SPICE* [31]. It aims the same task than CIDEr [19]. *Earth Mover Distance (EMD)* [32]. Although it is originally used to compare histograms, several authors use it to compare the image regions highlighted in their explanation to image regions highlighted by human judges, or in annotated images in XAI for VQA datasets [19]. *Rank correlation* [33]. It is a metric to transform two images in a list of pixels, and compare them. *Faithfulness score* [19] that checks the consistency between the visual explanation vectors obtained with a well-know explainer from the textual explanation, and the predicted answer from the VQA module. *Intersection of Union between* RoIs in the attention mask and the ground truth is used as a measure of visual explanation accuracy by the authors of [24]. *Failure case Analysis* is a manual qualitative process where the author presented cases where the answer predicted by the VQA model is different from the ground-truth, however, the answer was correct and the explanation was accurate [34].

5. Challenges and Future Work on VQA related to XAI

In this section, we want to explore the role of VQA related to XAI, not only about how we are going to explain VQA in the future, but how VQA could support XAI methodologies to improve users' trust on different AI models. VQA can play a key role in explanation comprehension and past literature on XAI has highlighted the need to personalise explanations to reduce the cognitive burden on the user [35]. Tailoring explanations to match the mental model of the user and the ability to interact with the user to provide clarifications, scrutinise and argue [36] are some of the approaches proposed to address personalisation. We find that VQA provides an opportunity to implement such interactions and improve comprehension and user satisfaction. The Alipour et al.'s work [26] introduces a SOTA architecture to solve the problem of VQA. Using the features from the images to apply a ResNet algorithm, authors obtain attention maps that point out the important part in the image to get an answer. However, the users can interact with the image, drawing the parts that they consider the most important. That new knowledge is included to execute the SOTA architecture, so the prediction is calculated again. As we can see there are some existing works that include interactivity in their explanations to improve them according to users' mental model. However, there is so much work left to do regarding this topic, for example using VQA to help users to understand the visual outputs generated by well-known explainers, like LIME, Grad-CAM, or Anchors. These outputs are explanations itself for AI models and tasks, but they might be confusing for users.

Example mock-up of how VQA can be used in post-explanation interactions is depicted in Figure 2. In this Figure we display an example of possible questions and the answers that could be obtained with VQA. On the left side, we show visual explanations obtained when applying a XAI technique (GradCAM, LIME, or Nearest Neighbours) to explain an artificial inteligence task (for example, image classification or text classification). Then, we treat the explanation to be the visual component. On the right side, we have examples of possible user questions to ask about the visual explanations, and the answers to be provided by VQA. The questions can be about the explanation, and even about the prediction or the visual input itself. A particular methodology that might be used is the one proposed by Kim et al. [37], in which they generate explanations for charts (of any kind) using VQA methods. This type of proposal could be useful to explain charts obtained with LIME, Anchors, etc. for tabular or text data. Moreover, such VQA model can be integrated with an existing interactive model that implements interactive-XAI [36]. There are many challenges in implementing such methods and they vary by the preferred approach: knowledge-intensive or knowledge-light.



Q: Can you describe the attention weights on this image for the "dog" prediction?

A: There is higher weight on the centre of the dog snout and eyes, a medium weight on the surrounding face and lower weight on the forehead

Q: What kind of a dog is it?

A: It is a Chow Chow.

Q: What are the other possible classification ?

A: The rest of the image contains a cat which is a possible prediction.

Q: What feature(s) most contributed to the mushroom classified as poisonous?

A: Odour has contributed the most for the mushroom to be poisonous

Q: What kind of mushroom is it?

A: The AI model predicts if a mushroom is poisonous or not. There is not enough information in this explanation to answer that question.

Q: What does this explanation tell about the query?

A: This explanation shows two examples that are similar to the query, one labelled "Normal" similar to the query, another similar example labelled "Abnormal".

Q: What is the possibility of the query being labelled "Abnormal"?

A: The AI model is not very confident about the "Normal" prediction, and the query is very similar to the "Abnormal" example, therefore there is a high possibility the query being "Abnormal".

Figure 2: VQA applied to support XAI. Example of possible questions and their answers that could be obtained with VQA

A knowledge-light approach to VQA will find the diversity and complexity of the visual explanations challenging. As seen in Figure 2, the three explainers generate significantly different explanations. However, the task can be simplified if the VQA model is implemented on a single data modality (image vs tabular) or a specific explainer (LIME vs GradCAM). Another key challenge is the lack of data for fine-tuning or training such VQA model. In a knowledge-light data-driven approach empirical evaluation should indicate how well the pre-training transfer to question answering of visual explanations, however, much effort is needed to create datasets that can be used for fine-tuning and training.

Regarding challenges related to knowledge-intensive methods, the largest one might be compiling a comprehensive database for VQA [38] that not only helps to increase the performance of VQA methods but also their explainability. Doing that, we could build extensive graphs to represent our knowledge and extract from them, more detailed and accurate explanations. Moreover, this task must be also helpful when we want to apply VQA to explain explanations generated by well-known explainers. These datasets will also be helpful when evaluating VQA to explain explanations themselves. However, we could propose to use GPT-3 to generate explanations from this kind of images. Moreover, there is only a limited amount of state-of-the-art techniques for knowledge-intensive VQA methods that are explainable. Although some authors claim that knowledge-intensive methods are transparent and can help users to understand the reasoning process carried out by the system [2], there is a lack of focused research. Therefore, more approaches are necessary, especially in evaluating the quality of explainability.

Another line of future research work could be methods that combine transformers-based architectures with knowledge bases. This could be an opportunity to enhance both proposals and overcome their weaknesses taking advantage of the strong features of each type of methodology.

Answers for questions that require additional knowledge-based reasoning and commonsense reasoning inference (see Section 3) could be improved this way.

Finally, we want to remark that there is not Case-Based Reasoning (CBR) methodologies applied to VQA or CBR applied to XAI for VQA. Therefore, exploring the role of CBR for this task could be interesting for the CBR community.

6. Conclusions

As with many other AI models, Visual Question Answering (VQA) models requires explanations for their target users to help them to understand how the model has predicted an answer for a specific question regarding the content in an image. In this work, we have studied some of the literature regarding VQA methods in general to comprehend the problem that we need to explain and the explainability of VQA methods. There are some promising approaches to implementing explainability into VQA methods, however, they are sparse often specific to the VQA method. We have found out that there are two key approaches to VQA: data-driven methods and knowledge-intensive methods. Former VQA methods that performed VQA as a pipeline of sub-tasks using CNN and RNN methods are considered transparent. More recent knowledge-light and intensive methods required ante-hoc or post-hoc explainability. In challenges for the future, we emphasise the possibility of using VQA to answer questions about explanations generated by well-known explainers. The addition of interactivity can also lead to enhanced user satisfaction in XAI systems since we can incorporate the user's mental model, allow disagreements and discover questions not answered by current XAI methods.

Acknowledgments

Supported by the Horizon 2020 Future and Emerging Technologies (FET) programme of the European Union through the iSee project (CHIST-ERA-19-XAI-008, PCI2020-120720-2). Funding in Spain by MCIN/AEI/10.13039/501100011033 and in the UK by EPSRC grant number EP/V061755/1.

References

[1] S. Shakya, et al., Analysis of artificial intelligence based image classification techniques, Journal of Innovative Image Processing (JIIP) 2 (2020) 44–54.

[2] Q. Wu, et al., Visual question answering: A survey of methods and datasets, Computer Vision and Image Understanding 163 (2017) 21–40.

[3] D. Gunning, D. Aha, Darpa's explainable artificial intelligence (xai) program, AI magazine 40 (2019) 44–58.

[4] A. B. Arrieta, et al., Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information fusion 58 (2020) 82–115.

[5] C. C. Aggarwal, et al., Recommender systems, volume 1, Springer, 2016.

[6] Y. Zhang, P. Tiňo, A. Leonardis, K. Tang, A survey on neural network interpretability, IEEE Transactions on Emerging Topics in Computational Intelligence 5 (2021) 726–742.

[7] J. Yang, J. Duan, S. Tran, Y. Xu, S. Chanda, L. Chen, B. Zeng, T. Chilimbi, J. Huang, Vision-language pre-training with triple contrastive learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 15671–15680.

[8] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, S. C. H. Hoi, Align before fuse: Vision and language representation learning with momentum distillation, Advances in neural information processing systems 34 (2021) 9694–9705.

[9] P. Wang, Q. Wu, C. Shen, A. Dick, A. Van Den Henge, Explicit knowledge-based reasoning for visual question answering, in: Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2017, pp. 1290–1296.

[10] S. Antol, et al., Vqa: Visual question answering, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 2425–2433.

[11] K. Basu, F. Shakerin, G. Gupta, Aqua: Asp-based visual question answering, in: Practical Aspects of Declarative Languages: 22nd International Symposium, PADL 2020, New Orleans, LA, USA, January 20–21, 2020, Proceedings 22, Springer, 2020, pp. 57–72.

[12] K. Marino, M. Rastegari, A. Farhadi, R. Mottaghi, Ok-vqa: A visual question answering benchmark requiring external knowledge, in: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition, 2019, pp. 3195–3204.

[13] A. v. d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, arXiv preprint arXiv:1807.03748 (2018).

[14] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 9729–9738.

[15] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, T. Duerig, Scaling up visual and vision-language representation learning with noisy text supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 4904–4916.

[16] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, J. Gao, Vinvl: Revisiting visual representations in vision-language models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 5579–5588.

[17] A. Agrawal, D. Batra, D. Parikh, A. Kembhavi, Don't just assume; look and answer: Overcoming priors for visual question answering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4971–4980.

[18] Q. Li, J. Fu, D. Yu, T. Mei, J. Luo, Tell-and-answer: Towards explainable visual question answering using attributes and captions, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 1338–1346.

[19] J. Wu, R. Mooney, Faithful multimodal explanation for visual question answering, in: Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, 2019, pp. 103–112.

[20] T. Stoikou, M. Lymperaiou, G. Stamou, Knowledge-based counterfactual queries for visual question answering, arXiv preprint arXiv:2303.02601 (2023).

[21] S. Ravi, A. Chinchure, L. Sigal, R. Liao, V. Shwartz, Vlc-bert: Visual question answering with contextualized commonsense knowledge, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 1155–1165.

[22] M. Sap, R. Le Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, Y. Choi, Atomic: An atlas of machine commonsense for if-then reasoning, in: Proceedings

of the AAAI conference on artificial intelligence, volume 33, 2019, pp. 3027–3035.

[23] B. X. Nguyen, T. Do, H. Tran, E. Tjiputra, Q. D. Tran, A. Nguyen, Coarse-to-fine reasoning for visual question answering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4558–4566.

[24] W. Tian, H. Li, Z.-Q. Zhao, Dual capsule attention mask network with mutual learning for visual question answering, in: Proceedings of the 29th International Conference on Computational Linguistics, 2022, pp. 5678–5688.

[25] R. Vaideeswaran, F. Gao, A. Mathur, G. Thattai, Towards reasoning-aware explainable vqa, arXiv preprint arXiv:2211.05190 (2022).

[26] K. Alipour, et al., A study on multimodal and interactive explanations for visual question answering, arXiv preprint arXiv:2003.00431 (2020).

[27] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[28] S. Banerjee, A. Lavie, Meteor: An automatic metric for mt evaluation with improved correlation with human judgments, in: ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005, pp. 65–72.

[29] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.

[30] R. Vedantam, C. Lawrence Zitnick, D. Parikh, Cider: Consensus-based image description evaluation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4566–4575.

[31] P. Anderson, B. Fernando, M. Johnson, S. Gould, Spice: Semantic propositional image caption evaluation, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14, Springer, 2016, pp. 382–398.

[32] O. Pele, M. Werman, A linear time histogram metric for improved sift matching, in: Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part III 10, Springer, 2008, pp. 495–508.

[33] A. Das, H. Agrawal, L. Zitnick, D. Parikh, D. Batra, Human attention in visual question answering: Do humans and deep networks look at the same regions?, Computer Vision and Image Understanding 163 (2017) 90–100.

[34] Y. Lin, et al., Revive: Regional visual representation matters in knowledge-based visual question answering, arXiv preprint arXiv:2206.01201 (2022).

[35] Q. V. Liao, D. Gruen, S. Miller, Questioning the ai: informing design practices for explainable ai user experiences, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020, pp. 1–15.

[36] P. Madumal, T. Miller, L. Sonenberg, F. Vetere, A grounded interaction protocol for explainable artificial intelligence, in: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, 2019, pp. 1033–1041.

[37] D. H. Kim, E. Hoque, M. Agrawala, Answering questions about charts and generating visual explanations, in: Proceedings of the 2020 CHI conference on human factors in computing systems, 2020, pp. 1–13.

[38] A. Cláudia Akemi Matsuki de Faria, et al., Visual question answering: A survey on techniques and common trends in recent literature, arXiv e-prints (2023) arXiv–2305.