Post-hoc XAI method for Visual Question Answering(VQA)

Muddamsetty, Satya Mahesh; B.Schmidt, Alina; Moeslund, Thomas B.

Published in:
International Conference on Pattern Recognitin and Artifical Intelligence (ICPRAI)-2024

Publication date:
2024

Post-hoc XAI method for Visual Question Answering(VQA)

Satya M. Muddamsetty[0000−0003−0935−4609], Alina B.Schmidt, and Thomas B. Moeslund[0000−0001−7584−5209]

Visual Analysis of Perception Laboratory (VAP), Aalborg University, Rendsburggade 14, 9000 Aalborg, Denmark
{smmu,tbm}@create.aau.dk { alina.schmidt}@centrica.com

Abstract. Visual Question Answering (VQA) systems, while advancing in intelligence, still face challenges in handling complex queries. Understanding the behavior of VQA models is crucial, especially in assessing their reliability in identifying relevant image regions for accurate responses. Although post-hoc explainable artificial intelligence (XAI) methods have demonstrated success in simpler tasks like image classification, their efficacy in more complex tasks like VQA remains unexplored. Moreover, the lack of a standardized evaluation method for the explanation heatmaps in VQA models raises questions about the trustworthiness of XAI methods. This paper proposes extending SIDU-XAI to VQA applications. It assesses visual explanations using the VQA-HAT dataset, which includes human-attention ground-truth maps, and the CLEVR-XAI dataset, comprising synthetic 3D scenes with objective ground-truth masks. Additionally, user studies were conducted to assess the impact of our post-hoc visual explanations on users' understanding of model decisions. Our experiments led to the conclusion that the proposed SIDU-VQA method surpasses state-of-the-art RISE-VQA methods in both quantitative and qualitative aspects across three experiments. However, post-hoc XAI methods fall short of entirely fulfilling the complex standards of human expectations, which highlights the critical need for continued research into XAI techniques appropriate for VQA applications.

Keywords: Visual Question Answering (VQA) · Explainable AI (XAI) · post-hoc XAI · SIDU · Deep Learning.

1 Introduction

The recent and continuing success of deep learning has meant that deep neural networks are deployed for automated decision-making more often than ever before. It has been applied to several computer vision tasks, such as object recognition [15], image classification [18], and image captioning [10]. Over the past few years, the task of visual question answering (VQA) has been widely investigated, leveraging the advancements in natural language processing (NLP)

and computer vision (CV). The main goal of a visual question answering (VQA) model is to respond to a natural language query related to the content of an image or any of the objects present in it [25]. Visual question answering is comparatively more challenging than image captioning since it frequently requires information that is not included in the image. The additional information needed can vary from common sense to encyclopedic knowledge about a specific element in the image [25].

VQA systems [26] are trained to simply fit the answer distribution using question and visual features, achieving significant performance levels on simple visual questions. However, these systems often exhibit poor explanatory capabilities and take shortcuts by focusing solely on simple visual concepts or question priors instead of finding the right answer for the correct reasons [14]. This problem becomes increasingly severe when the questions require more complex reasoning and commonsense knowledge. In reality, this bias is not easily noticeable because users frequently pose identical questions about the picture or the objects that are shown, and they are almost certainly aware of the right answer. For more complicated questions or when the user lacks the knowledge and experience of the questions or image content (for instance, in, the medical domain), it will be impossible to capture the behavior of the VQA system and determine whether it is biased. Furthermore, the available datasets are usually unbalanced for certain types of questions. This is because the dataset annotators tend to generate questions about objects detectable in the image, which causes the dataset to suffer from the so-called "visual priming bias." [7]. Moreover, for the deployed VQA systems, users need to understand how the VQA model makes decisions to trust them and apply them most effectively [25].

Therefore, it is very important to interpret the results of these systems and ascertain what caused the model to provide an answer based on the image-question pair. Nonetheless, it can be very challenging to establish confidence with deep learning-based VQA models as they are notably opaque, challenging to understand, and frequently exhibit unexpected failure modes. Explainable AI (XAI) is an emerging discipline that helps to understand how AI models make decisions, and to find possible weaknesses that can be improved upon. Post-hoc XAI methods have proven capabilities, allowing users to look inside the black box by demonstrating which features of the input are significant in the system's decision-making process. Several other post-hoc XAI methods have been proposed in the literature [2], that provide visual explanations in the form of saliency maps. GradCAM, [22] for example, analyses the gradient space to find the visual regions that most affect the decision. However, such methods have mainly been tested and evaluated on image classification-based black box models. Compared with single-object image classification, the VQA task is more difficult because it requires a more selective and spatially grounded setup for evaluating visual explanations. Furthermore, not every object in the image has predictive value in VQA [5]. The question influences the explanation since it de facto chooses which objects are significant for a given prediction (i.e., the heatmap depends on more than just the content of the image), whereas, in

typical image classification tasks, each image has a single relevant object that needs to be identified to reach a final classification decision. The effectiveness of post-hoc XAI approaches, however, is still uncertain in respect to VQA models. Moreover, there are a limited number of papers in the literature that focus on the interpretability of VQA models. Nor is there an established protocol for evaluating explainable VQA methods in the existing literature. Therefore, in this paper, we aim to investigate how post-hoc XAI methods can be useful in understanding VQA model decisions and answering different types of questions based on image content. In particular, we extend the post-hoc XAI method SIDU [17] for the VQA model that uses CNN and LSTM models. Our extended VQA-XAI method takes the last convolution layer of the CNN model to generate the feature image masks. Feature importance weights (similarity difference and uniqueness weights) are subsequently computed using the predictions of the VQA (CNN + LSTM) model for each feature image mask. We propose a thorough evaluation framework focusing on 'human attention' and 'objectness', utilizing two different VQA-HAT [9] and Compositional Language and Elementary Visual Reasoning Diagnostics (CLEVR-XAI) [6] datasets that are devoted to assessing the VQA model explanations. Furthermore, we conducted a user study to evaluate the usefulness of explanations for helping users understand how a VQA system handles different types of questions. The main contributions are summarized below:

1. For the first time, we expand the SIDU XAI method to the Visual Question Answering (VQA) application, which involves a combination of both image and text models (CNN + LSTM). Specifically, we utilized the final convolution layer of the VQA model's CNN to generate feature activation image masks. The prediction of these masks is computed using the VQA model, which has inputs of both images and questions, ultimately determining the final SIDU weights for generating visual explanations.
2. We showcase the efficacy of the SIDU-VQA XAI method and conduct a comparative analysis utilising the state-of-the-art RISE-VQA explanation method. Our evaluation is carried out on the VQA Human Attention dataset and the CLEVR XAI dataset, considering both human attention and objectness aspects, through quantitative and qualitative assessments. Across both datasets, SIDU-VQA outperforms RISE-VQA in overall performance.
3. We conducted a novel user study aimed at comprehending the post-hoc-generated visual explanation maps for VQA models. The main objective of this user study is to assess the usefulness of heatmap-based visual explanations to understand the VQA model's decision. Our subjective evaluation reveals that SIDU-VQA visual explanations are more useful than RISE-VQA in understanding model decisions.

The remainder of the paper is organized as follows: Section 2 presents state-of-the-art XAI methods for VQA and XAI evaluation methods. The SIDU-XAI-VQA method is explained in Section 3. Section 4 presents the experimental evaluations and results, and Section 5 concludes the study and discusses future work.

2 Related work

In this section, we present some of the state-of-the-art Visual Question Answering (VQA) visual explanation methods in relation to various evaluation methods. The literature for each direction is presented in the following subsections:

2.1 XAI methods for VQA

Current research within the field of VQA has produced several XAI methods intended specifically for the VQA task. One XAI approach explicitly designed for VQA includes building interpretable neural networks, like the Neural Module Network (NMN) [4] [11] architectures, which parse questions into linguistic parts that can be modeled as subtasks, then combine pre-defined modules that can solve each subtask to create a deep network specific to each question. However, this approach limits the kinds of networks that can be used and is not useful for explaining decisions made by state-of-the-art VQA networks. Several state-of-the-art VQA networks use attention mechanisms. Attention mechanisms allow the model to focus on a subset of the visual representation and can create an image-based visualization of the attention weights that can be repurposed to deliver insights into the decision of the model without using any specific explainable method. The authors in [25, 23] use these "attention maps" in order to visualise the model's behaviour and to both evaluate and explain how the model works. However, the explainable quality of the maps themselves is rarely evaluated in quantitative terms. For example, some modules in End-to-End Module Networks (N2NMNs) by Hu et al. [11] output 'attention maps', but the network is only measured/evaluated as regards the accuracy of the predictions. Similarly, the Stacked Attention Network by Yang et al. [27] produces two attention maps corresponding to the two layers of attention mechanisms. In both cases, no evaluation of the heatmaps themselves was conducted by the authors. Another example is the Hierarchical Co-Attention (HieCoAtt) architecture by Lu et al. [16], in which the model outputs co-attention maps for both the image and at a word-level for each image-question pair. While the attention maps were not evaluated by the authors. The authors in [3] proposed a combined bottom-up and top-down attention mechanism to calculate attention at the object level. This model was further upgraded and fine-tuned to win the VQA Challenge 2018. Another approach is to take post-hoc XAI methods made for different tasks involving CNNs and and apply them to the VQA task. One of these is Grad-CAM [22]. Grad-CAM is a gradient-based algorithm that uses back-propagation to track a network's activations from the output back to the input in order to create a saliency map. The saliency maps created by Grad-CAM were evaluated on the VQA-Human Attention dataset (VQA-HAT). However, Grad-CAM is restricted by the problem of gradient saturation, which results in reduced back-propagating gradients. Due to that, the quality of visualizations is negatively impacted [24]. In contrast, we demonstrate a novel "gradient-free" post-hoc visual explanation approach, SIDU–VQA for interpreting VQA models.

2.2 XAI evaluation for VQA

Evaluating the efficiency of VQA-XAI methods requires a variety of strategies, each with unique benefits and limits. To evaluate the visual explanations, two main factors are considered in the literature: faithfulness and human interpretability [21]. Faithfulness requires that the explanations should faithfully represent the decisions of the model. One metric for evaluating faithfulness is Insertion and Deletion [19]. These processes work in complementary ways by either removing or adding pixels to the image in the order of the pixels that the saliency map deems most important. The probability for the predicted class changes based upon how many pixels have been added or removed. The more faithful the generated saliency map is, the faster the prediction will change when you remove or add the most important pixels. Meanwhile, the authors in [20] proposed a pixel perturbation analysis to evaluate faithfulness. These faithfulness evaluations were mostly validated on image classification tasks. Furthermore, human interpretability requirements dictate that the saliency. Human interpretability declares that the saliency maps should be interpretable and intuitive to humans. The majority of XAI methods that are applied to VQA are evaluated along these lines. The authors in [22] use human-generated saliency maps as ground truth for evaluating machine-generated equivalents. Indeed, the popular Visual Question Answering: Human Attention (VQA-HAT) dataset [9] takes this approach. The dataset is based on the Visual Question Answering (VQA) 1.0 dataset and was created by having individuals manually mask the parts of natural images they found useful for answering questions. The masks can then be compared to machine-generated ones using rank correlation. In this way, the XAI methods are tested to ascertain whether their output matches what humans require when answering the questions. However, one downside to using human-generated masks is that humans are subjective, not particularly precise, and do not always agree. Thus, the authors proposed a CLEVR-XAI [6] dataset, which is a synthetic dataset with images of nondescript 3D scenes containing random objects followed by strategic questions about the contents of these scenes that have been designed to only involve specific objects within each scene. In this way, the exact location of the objects required to answer the questions can be isolated, and machine-generated saliency maps can be tested against these ground truths in order to measure if they focus on the parts of the images containing the relevant objects, the same strategy humans would employ. Alternatively, some of the methods are evaluated by conducting a user study to assess the quality of explanations [16]. However, the literature on interpreting and explaining VQA system outcomes is relatively sparse, and most interpretable VQA models employ an attention mechanism. Thus, it is notable that existing approaches have primarily concentrated on appraising VQA model performance, often foregoing a comprehensive evaluation of the effectiveness of explanations [16]. Additionally, there is a lack of standardized procedures for assessing explainable VQA methods in the existing literature. While some methods have undertaken assessments using human attention datasets alone, the effectiveness of VQA-XAI methods cannot be adequately assessed when based on a single aspect. Our approach stands out

by conducting both quantitative and qualitative evaluations, considering human attention, objectness, and usefulness as three distinct aspects phenomena, with which to comprehensively evaluate the effectiveness of XAI methods in VQA applications.

3 Explainable VQA method (SIDU-VQA)

In this section, we outline the specific method employed for visual question-answering explanation. Visual question-answering aims to predict an answer based on a given question and an image. The main goal of visual question-answering is to predict an answer from a given question and an image. Previous work, such as SIDU, demonstrated its efficacy in providing insightful visual explanations for comprehending feed-forward neural network classification decisions in the image domain [17]. Building upon this, we extend the application of SIDU to the Visual Question Answering (VQA) model, denoted as SIDU-VQA. Unlike a classification model, the VQA model processes both images and questions simultaneously. SIDU-VQA is crafted to elucidate the predicted output of the black-box VQA model (CNN+LSTM) employed for answering questions based on provided input images.

Initially, we generate feature image masks by extracting the last convolution layer of the VQA model (CNN + LSTM), denoted as $\mathcal{F}$. The last convolution layers, with dimensions $n \times n \times N$, where $'n'$ is the size of the convolution layer, and $'N'$ is the total number of feature activations $\mathbf{f}$ for class c, are represented as $\mathbf{f}^c = [f^c1, ...f^cN]$. Each feature activation map f^ci is then converted into a binary mask B^ci through thresholding each value, defined as:

$$B^c_{i=1..N} = f^c_{i=1..N} > \tau \tag{1}$$

Here, τ represents the threshold. The binary mask B^ci is up-sampled using bilinear interpolation for a given input image I with dimensions $Width \times Height$. Subsequently, point-wise multiplication is performed between the feature activation mask (up-sampled binary mask) M^ci and the input image I to calculate the feature activation image mask A^i_c, represented as:

$$A^c_i = F(I \odot M^c_i), \tag{2}$$

Subsequently, the feature activation image mask and the input question are fed into the VQA model to compute the prediction vector P^c_i. Similarity differences and uniqueness weights for each mask concerning a predicted answer c are computed, and the weights of each mask are combined into a final map, illustrating the explanation of the prediction.

The similarity difference SD^c_i measure between the prediction vector of the original input image I ($P^cvqaorg$) and the i^{th} feature activation image mask prediction (P^ci) is given by:

$$SD^c_i = \exp\left(\frac{-\|P^c_{vqaorg} - P^c_i\|}{2\sigma^2}\right) \tag{3}$$

where σ is a controlling parameter and the uniqueness U_i^c measure is defined as:

$$U_i^c = \sum_{j=1}^{N} \|P_i^c - P_j^c\|, \qquad i = 1, 2, ..., N \tag{4}$$

The SIDU weights are computed as the product of the similarity differences and uniqueness measures:

$$w_i^c = SD_i^c \cdot U_i^c, \tag{5}$$

Finally, to obtain the visual explanation (saliency map) of the predicted output answer c of a VQA model F, a weighted sum is performed between the feature activation mask $M^c i$ and the corresponding feature importance weights $W^c i$. The visual explanation of the predicted class c is given by:

$$S_{VQA}^c = \frac{1}{N} \sum_{i=1}^{N} w_i^c \cdot M_i^c \tag{6}$$

4 Experimental Evaluation

In this section, we assess the performance of the SIDU-VQA methods through a series of comprehensive experiments. Our evaluation focuses on studying the effectiveness of the visual explanation of model predictions in response to the input question. To achieve this, we leverage two datasets—VQA-HAT (Visual Question Answering Human Attention) [9] and CLEVR-XAI [6]—which are well-suited for evaluating XAI methods in VQA applications. Additionally, we conducted a user-experimental study, in order to assess the usefulness of visual explanations.

4.1 Evaluating VQA-XAI with Human Attention

Visual questions focus on specific characteristics of an image, including background elements and underlying context. This implies that an explicit or implicit attention strategy might help a VQA model to appropriately respond to a query. To ascertain the alignment between visual explanations provided for a given question and human attention based on image content, we delve into attention analysis in the context of VQA. In particular, we are especially curious about the following queries: a) Which areas of an image do individuals naturally choose to focus on when seeking answers to questions related to that image? b) Do deep VQA models, when given post-hoc explanations, target the same regions as humans?

Hence, our selection is the Visual Question Answering Human Attention dataset (VQA-HAT) [8] which has been specially designed to understand human behavior on the VQA application. VQA-HAT comprises ground truth masks for 'human attention' related to a subset of the Visual Question Answering (VQA) 1.0 dataset [1]. The dataset consists of both the training set with 58,475 attention masks and the validation set with 4122 masks in total, 3 for each of the 1374

image/question pairs in the dataset. An innovative interface inspired by multiple games is designed to collect human attention maps, revealing where individuals focus in order to answer questions within the extensive VQA dataset [8].

To quantitatively evaluate the effectiveness of post-hoc VQA-XAI methods' explanations, we selected a validation set comprising 4,122 masks. For comparison with other methods on the VQA-HAT dataset, we chose the VQA network HieCoAtt [16], which exhibits an overall performance with a 60% accuracy rate on the VQAv1 datasett [1]. RISE, a state-of-the-art post-hoc method, was selected for comparison due to its superior performance over gradient-based methods [19]. We assessed visual explanation maps generated by SIDU-VQA and the state-of-the-art method RISE-VQA against human attention, both qualitatively (via visualizations) and quantitatively (via rank-order correlation). The rank-order correlation averaged across all image-question pairs in the validation set is summarized in Table 1. It is clear that, SIDU-VQA achieved a rank correlation score of 0.234, while RISE-VQA achieved a rank correlation score of 0.0067. These results affirm that the visual explanations provided by SIDU-VQA outperform the state-of-the-art RISE-VQA, as depicted in Figure 1. From the figure, it is evident that SIDU-XAI maps are positively aligned with human attention maps. In contrast, RISE-XAI fails to align with human attention.

Model	Rank-correlation↑
SIDU-VQA	**0.234**
RISE-VQA [19]	0.0067

Table 1. Mean rank-correlation coefficients results (higher is better) for the SIDU method compared with state-of-the-art method RISE-VQA. All models were evaluated on the validation split of the VQA-HAT dataset.

4.2 Evaluating on CLEVR-XAI

In the above section, we evaluated the XAI-VQA method explanation maps using human attention, which can demonstrate how effectively the visual explanation of VQA models is correlated with human attention maps. In this section, we evaluate the objectness of the VQA-XAI method on the recently released CLEVR-XAI dataset [6] which is based on the compositional language and the Elementary Visual Reasoning Diagnostics (CLEVR) dataset [12]. CLEVR is a synthetic VQA task that was designed to diagnose the reasoning abilities of VQA models by avoiding the biases present in real-world human-annotated datasets, and allowing full control of the data generation pipeline. The dataset consists of randomly generated synthetic 3D scenes of a nondescript grey environment containing several spheres, cubes, and/or cylinders of different sizes and made of metal or rubber. For each image, questions about one or more objects and pixel-level ground truth masks, corresponding to the objects that are being asked about, were generated using a functional program. The main advantage of the

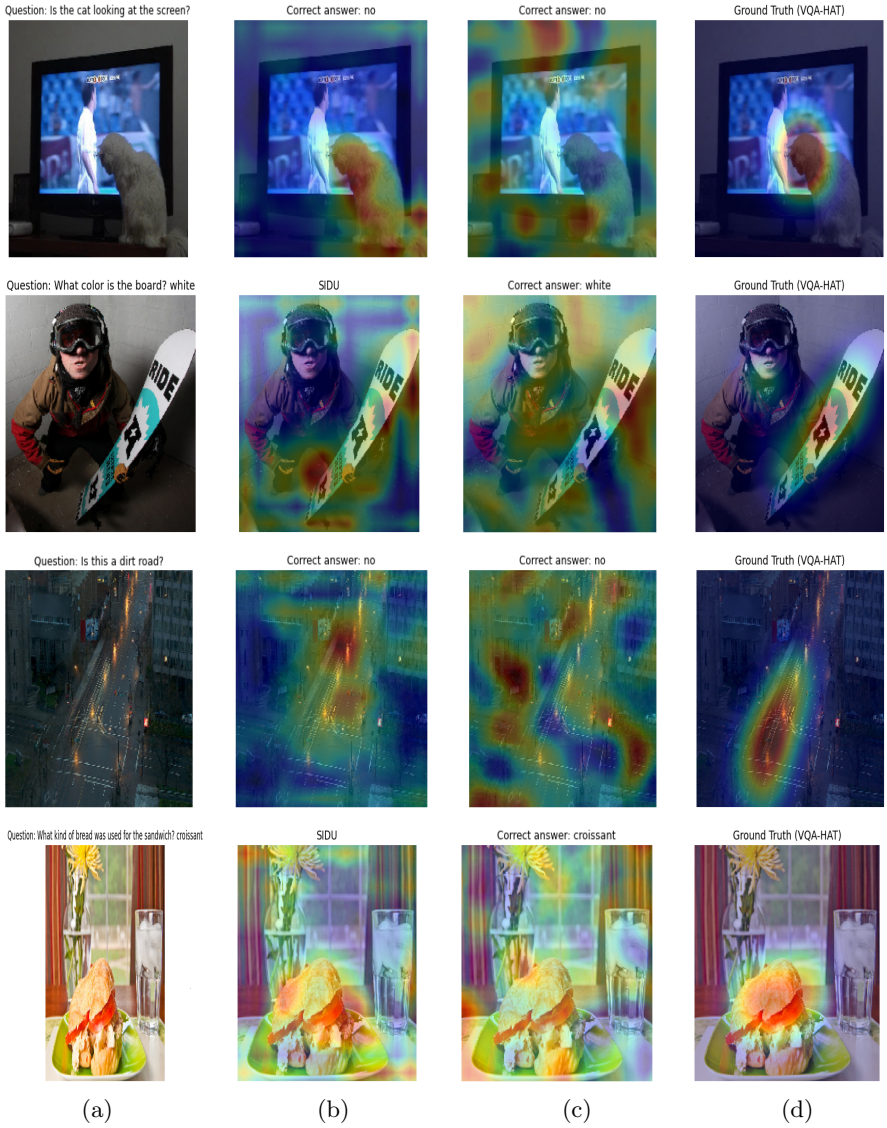

Fig. 1. Example of visual explanations generated by SIDU-VQA and RISE-VQA for the VQA model trained on the VQA V1 dataset.(a)original input image and question (b) SIDU-VQA. (c) RISE-VQA(d) Human attention mask

dataset is that it provides a bias-free, selective, controlled, and realistic testbed for the evaluation of VQA model explanations. This dataset enables us to assess the method's reliability by consistently emphasizing the same object across various questions associated with that specific object. For this assessment, we select the VQA model (CNN+LSTM), pre-trained on the CLEVR dataset with

an accuracy of 75% [13]. RISE, a state-of-the-art post-hoc method, is chosen for comparison. The primary reason behind this selection is that what exhibits superior performance compared to gradient-based methods [19]. The visual explanation maps were generated with the SIDU-VQA and RISE-VQA methods using the VQA model for randomly selected 1000 image/question pairs from the CLEVR-XAI dataset. We chose the AUC metric to evaluate the performance of the XAI methods using CLEVR-XAI ground truth masks. Table 2 summarizes the mean AUC results obtained by the XAI-VQA method. From the table, we can observe that SIDU-VQA outperforms RISE-VQA with a mean AUC of 0.840 on the CLEVR-XAI dataset, which is 20 % more than RISE-VQA. Observing the heatmap in Figure 2, it becomes apparent that the SIDU-VQA method accurately localizes precise objects with high importance, aligning closely with the ground truth for the provided input question. In contrast, the RISE-VQA method struggles to pinpoint the precise object for the input question, often localizing incorrect objects. In summary, the RISE-VQA method tends to provide misleading indications regarding the correct object, even when the model predicts the accurate answer.

Methods	Mean AUC ↑
SIDU-VQA	**0.840**
RISE-VQA	0.6107

Table 2. Mean area under the curve (AUC) results (higher is better) for the SIDU-VQA method compared RISE-VQA on CLEVR-XAI dataset.

4.3 Evaluating the usefulness of explanation: User-study Evaluation

The preceding experimental results establish the actual effectiveness of visual explanation maps for VQA applications. However, we also see that users are often misled by current heatmap-based visualizations that point to relevant regions, despite the model producing a correct answer. Therefore, we conducted a user study to evaluate human comprehension of the Visual Question Answering (VQA) model's decisions by utilizing visual explanations. The main objective of this user study is to assess the usefulness of heatmap-based visual explanations to understand the VQA model's decision. The visual explanation is held to be useful if it points to the relevant region for the VQA model's prediction.

To conduct these experiments, three human subjects participated in assessing the VQA-XAI visual explanations. They were presented with an image-question (IQ) pair and a heatmap and tasked with determining if an explainable AI (XAI) method could generate the correct explanation. One hundred images were randomly selected from the VQAv1 dataset. The human subjects received samples and visual explanations for two distinct post-hoc XAI methods—SIDU-VQA and RISE-VQA, explaining the VQA model's decision for the provided input

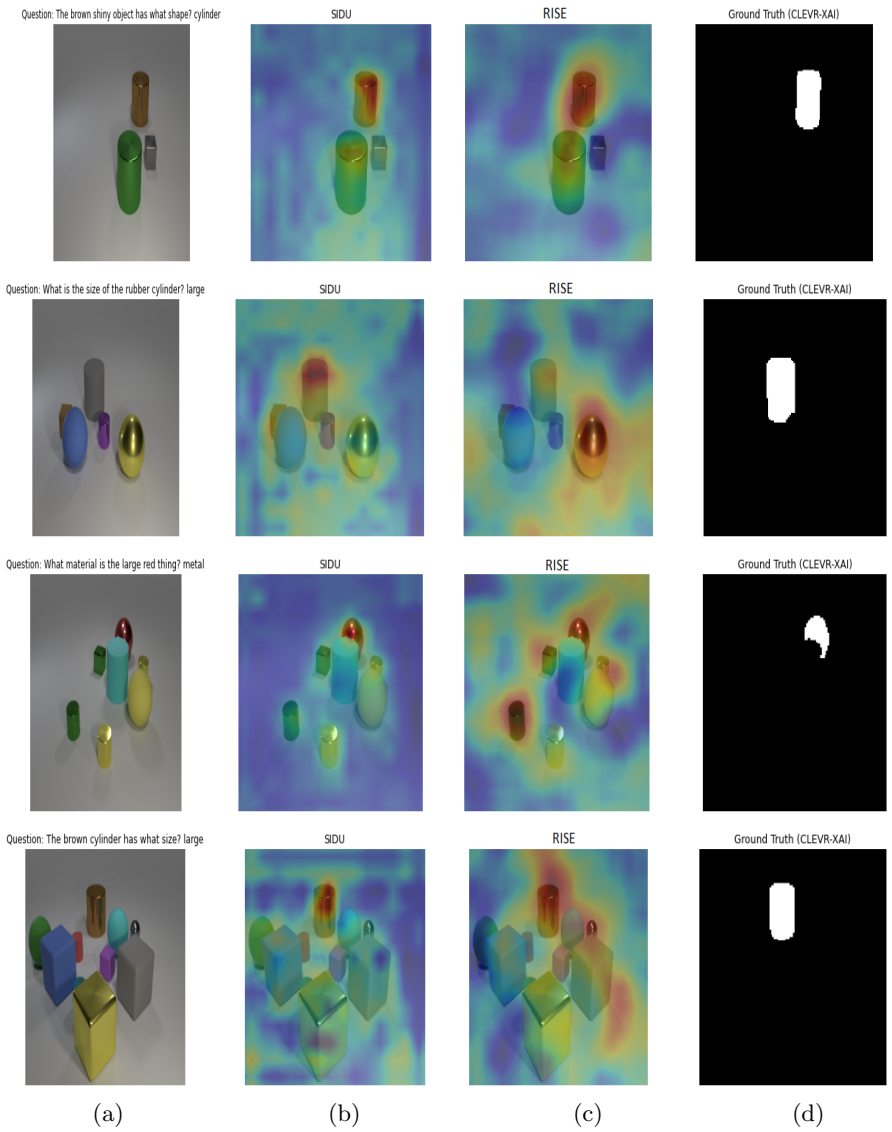

Fig. 2. Example of visual explanations generated by SIDU-VQA and RISE-VQA for the VQA model trained on CLEVR synthetic dataset and groundtruth object masks. (a)original input image and question (b) SIDU-VQA. (c) RISE-VQA (d) Ground truth object mask

questions. The explanation methods used the pre-trained VQA model as described in [16]. To eliminate bias, these models were anonymized and labeled as VQA-XAI Method I and VQA-XAI Method II. Participants were asked two

questions: Is the visual explanation useful in understanding the model's decision for the given input question? Which visual explanation is superior?

The users were asked to assess the methods individually when providing the answers. For the questions, options included 'VQA-XAI Method 1,' 'VQA-XAI Method 2,' 'both,' and 'None of them.' The users were asked to choose one of the options for the given test samples. Subsequently, the data collected from users underwent a majority voting strategy, determining the outcome for each image. We then calculated the percentage of each choice.

Analyzing the outcomes, we note that 44% of respondents favored SIDU-VQA explanations as useful, while only 6% opted for RISE-VQA explanations are useful. This indicates that visual explanations based on SIDU-VQA are more beneficial in comprehending the decisions of the VQA model compared to those based on RISE-VQA. Notably, 41% of participants chose 'None of them,' suggesting that heatmap-based visual explanations have limited utility for some users and may potentially mislead. To understand this decrease in usefulness, it is crucial to consider the question's complexity. When questions are straightforward, humans find it easier to grasp the model's decision, while in more complex scenarios, visual explanations alone may fall short. This discrepancy could be attributed to the prevalent use of unbalanced datasets in training current state-of-the-art VQA models, where the majority of annotations involve simple questions. The subjective evaluation leads to the conclusion that relying solely on visual explanations is insufficient for understanding the VQA model's output for the given input question.

5 Conclusion and Future work

In this paper, we have expanded the applicability of the post-hoc explainable AI (XAI) method SIDU to Visual Question Answering (VQA) scenarios, aiming to elucidate the model decisions for given input questions. The effectiveness of these expanded post-hoc XAI methods for VQA is comprehensively evaluated, with a specific emphasis on human attention, objectness, and usefulness. This evaluation utilizes two dedicated datasets designed for assessing visual explanations in VQA applications: one containing human-generated saliency maps (VQA-HAT) and another employing the CLEVR synthetic dataset with objective ground truth object masks. The findings suggest that SIDU-VQA outperforms other methods, demonstrating superior mean rank correlation and mean AUC metrics.

Moreover, a user study was conducted to gauge the usefulness of visual explanations in comprehending VQA model decisions. The study suggests that post-hoc XAI methods offer limited insight into understanding the model's decision, emphasizing the ongoing need for improved explanation methods in VQA applications to enhance human comprehensibility. The interpretability of VQA models is still in its early stages. Therefore, as a prospective avenue, we aim to enhance post-hoc XAI methods by incorporating textual explanations alongside visual explanations to guide VQA systems. While visual explanations highlight

image segments contributing most to the answer, textual explanations provide richer information, such as detailed attributes, relationships, or common-sense knowledge not necessarily evident in the image. Furthermore, we also aim to extend our novel evaluation framework to compare post-hoc XAI methods with attention mechanism VQA methods.

References

1. Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C.L., Parikh, D., Batra, D.: VQA: Visual Question Answering: www.visualqa.org. International Journal of Computer Vision **123**(1) (2017). https://doi.org/10.1007/s11263-016-0966-6
2. Alicioglu, G., Sun, B.: A survey of visual analytics for explainable artificial intelligence methods. Computers & Graphics **102**, 502–520 (2022)
3. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6077–6086 (2018)
4. Andreas, J., Rohrbach, M., Darrell, T., Klein, D.: Learning to Compose Neural Networks for Question Answering. In: Association for Computational Linguistics. pp. 1545–1554 (1 2016). https://doi.org/10.48550/arxiv.1601.01705, http://arxiv.org/abs/1601.01705
5. Arras, L., Osman, A., Samek, W.: Ground truth evaluation of neural network explanations with clevr-xai. arXiv preprint arXiv:2003.07258 (2020)
6. Arras, L., Osman, A., Samek, W.: CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations. Information Fusion **81** (2022). https://doi.org/10.1016/j.inffus.2021.11.008
7. Boukhers, Z., Hartmann, T., Jürjens, J.: Coin: Counterfactual image generation for visual question answering interpretation. Sensors **22**(6), 2245 (2022)
8. Das, A., Agrawal, H., Zitnick, C.L., Parikh, D., Batra, D.: Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions? In: Conference on Empirical Methods in Natural Language Processing (EMNLP) (2016)
9. Das, A., Agrawal, H., Zitnick, L., Parikh, D., Batra, D.: Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions? Computer Vision and Image Understanding **163**, 90–100 (10 2017). https://doi.org/10.1016/j.cviu.2017.10.001
10. Ghandi, T., Pourreza, H., Mahyar, H.: Deep learning approaches on image captioning: A review. ACM Computing Surveys **56**(3), 1–39 (2023)
11. Hu, R., Andreas, J., Rohrbach, M., Darrell, T., Saenko, K.: Learning to Reason: End-to-End Module Networks for Visual Question Answering. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 804–813 (4 2017). https://doi.org/10.48550/arxiv.1704.05526, http://arxiv.org/abs/1704.05526
12. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C.L., Girshick, R.: CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1988–1997 (12 2017). https://doi.org/10.48550/arxiv.1612.06890, https://arxiv.org/abs/1612.06890

13. Johnson, J., Hariharan, B., van der Maaten, L., Hoffman, J., Fei-Fei, L., Zitnick, C.L., Girshick, R.: Inferring and executing programs for visual reasoning. In: ICCV (2017)
14. Joshi, G., Walambe, R., Kotecha, K.: A review on explainability in multimodal deep neural nets. IEEE Access **9**, 59800–59821 (2021)
15. Liu, Y., Sun, P., Wergeles, N., Shang, Y.: A survey and performance evaluation of deep learning methods for small object detection. Expert Systems with Applications **172**, 114602 (2021)
16. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. In: Advances in Neural Information Processing Systems (2016). https://doi.org/https://doi.org/10.48550/arXiv.1606.00061
17. Muddamsetty, S.M., Jahromi, M.N., Ciontos, A.E., Fenoy, L.M., Moeslund, T.B.: Visual explanation of black-box model: Similarity Difference and Uniqueness (SIDU) method. Pattern Recognition **127**, 108604 (7 2022). https://doi.org/10.1016/j.patcog.2022.108604
18. Obaid, K.B., Zeebaree, S., Ahmed, O.M., et al.: Deep learning models based on image classification: a review. International Journal of Science and Business **4**(11), 75–81 (2020)
19. Petsiuk, V., Das, A., Saenko, K.: RISE: Randomized Input Sampling for Explanation of Black-box Models. In: Proceedings of the British Machine Vision Conference (BMVC) (6 2018). https://doi.org/10.48550/arxiv.1806.07421, http://arxiv.org/abs/1806.07421
20. Samek, W., Binder, A., Montavon, G., Lapuschkin, S., Müller, K.R.: Evaluating the visualization of what a deep neural network has learned. IEEE transactions on neural networks and learning systems **28**(11), 2660–2673 (2016)
21. Samek, W., Montavon, G., Lapuschkin, S., Anders, C.J., Müller, K.R.: Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications. Proceedings of the IEEE **109**(3), 247–278 (3 2021). https://doi.org/10.1109/JPROC.2021.3060483
22. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 618–626 (10 2017). https://doi.org/10.1007/s11263-019-01228-7, http://arxiv.org/abs/1610.02391 http://dx.doi.org/10.1007/s11263-019-01228-7
23. Srivastava, Y., Murali, V., Dubey, S.R., Mukherjee, S.: Visual Question Answering using Deep Learning: A Survey and Performance Analysis. In: Computer Vision and Image Processing (8 2021), http://arxiv.org/abs/1909.01860
24. Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., Hu, X.: Score-cam: Score-weighted visual explanations for convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 24–25 (2020)
25. Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A., van den Hengel, A.: Visual question answering: A survey of methods and datasets. Computer Vision and Image Understanding **163** (2017). https://doi.org/10.1016/j.cviu.2017.05.001
26. Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A., Van Den Hengel, A.: Visual question answering: A survey of methods and datasets. Computer Vision and Image Understanding **163**, 21–40 (2017)
27. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked Attention Networks for Image Question Answering. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21–29 (11 2015). https://doi.org/10.48550/arxiv.1511.02274, http://arxiv.org/abs/1511.02274