ARL-TR-9817 • Oct 2023

Anomaly Detection with Visual Question Answering

by Stephanie M Lukin and Rahul Sharma

NOTICES

Disclaimers

ARL-TR-9817 ● Oct 2023

Anomaly Detection with Visual Question Answering

by Stephanie M Lukin
DEVCOM Army Research Laboratory

Rahul Sharma
University of Maryland, College Park

REPORT DOCUMENTATION PAGE			*Form Approved* *OMB No. 0704-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED (From - To)**
October 2023	Technical Report	May 30–August 18, 2023

4. TITLE AND SUBTITLE
Anomaly Detection with Visual Question Answering

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Stephanie M Lukin and Rahul Sharma

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
DEVCOM Army Research Laboratory
ATTN: FCDD-RLA-IC
2800 Powder Mill Rd, Adelphi, MD 20783

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TR-9817

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT A. Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
Contact author email: <stephanie.m.lukin.civ@army.mil>

14. ABSTRACT
Anomaly detection is critical for many different use-cases, such as identifying safety hazards to potentially prevent disasters. Developing the capability for a human-robot team to ask targeted questions would be critical to quickly identify a violation of protocol and then quickly take action to rectify the situation. In this report, we experiment with how visual question answering algorithms can be used with a set of carefully constructed questions to detect anomalies in a virtual makerspace and a real-world alleyway. Our exploratory results show improvement over a random baseline and we discuss challenges for future work.

15. SUBJECT TERMS
anomaly detection, experimental design, visual question answering, Military Information Sciences

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
					Stephanie M Lukin
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER (Include area code)**
Unclassified	Unclassified	Unclassified	UU	19	310-448-5396

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures

List of Tables

1. Introduction

Human-guided robotic exploration can be useful for gathering information at remote locations, especially those that might be too risky, inhospitable, or inaccessible for humans due to hazardous conditions or natural disasters.[1,2] Site conditions may be unknown and continuously changing; therefore, having a robot visually monitor and report on the evolving scenario would help a human partner decide what action the robot should take next or what to investigate further. A robot with capabilities for automatic detection when a scenario is in violation of safety protocols or expectations may be able to support this goal.

The detection of anomalies has been explored in prior works at the individual entity level (e.g., a hole in a piece of fabric), at the scene level pertaining to a specific task (e.g., obstacle avoidance for self-driving cars), and at the events level (e.g., unexpected movement over the course of the video).[3,4] We situate our work within the scene level, and incorporate an interactive element for anomaly detection in human-robot teams.

In this report, we design a paradigm for visual scene analysis centering around safety protocols and subsequent detection of violations. We gather background knowledge about the expectations of a particular domain (makerspace safety protocols) and then ask a system questions about snapshots from the environment to assess if it is anomalous. Our approach applies recent advances in visual question answering (VQA) algorithms to a focused inquiry of visual sensory anomaly detection. We identify two sets of visual stimuli for our study: a virtual makerspace, which we manipulate to violate our curated safety protocols; and a real-world alleyway, which exhibits anomalous configurations and properties. The two domains, questions, and VQA answers are shown in Fig. 1.

In this work, we ask the following research question: *Can asking targeted questions identify and assess visual anomalies more accurately than requesting a generic description of the image (e.g., from an image captioning algorithm?)* Our contributions are as follows: 1) We design an anomaly categorization from safety protocols to guide both design of environments and the VQA algorithm's line of questioning, and 2) We assess how well state-of-the-art VQA performs in atypical and out-of-domain environments. In our proof-of-concept experiments, our VQA-targeted questioning approach is able to achieve 70% accuracy on anomalies in a makerspace image compared to 50% randomness and 0% accuracy using caption generation and description-only baselines. We discuss the potential in asking questions to assess anomalies, and describe future work to increase the scale of our testing.

Fig. 1 Two visual stimuli: a virtual makerspace (left) and a real-world alley (right). The VQA is asked the question "Is anything obstructing the ground around the fire extinguisher?" and it responds correctly to both stimuli "yes" and "no," respectively.

2. Visual Question Answering

VQA is an open-ended task where, given an image and a natural language question, the system provides an accurate natural language answer.[5] The task is designed to closely "reflect the challenge of general image understanding"[6] in that the VQA models do not know the question that will be asked of an image until run time; therefore, any number of possible questions may be applicable for each image. Questions that comprise VQA benchmarks take on many formulations: yes-no questions (e.g., "is..." or "are..."), counting questions (e.g., "how many..."), and identification of objects or people (e.g., "what kind..." or "who..."). There are several variants to the VQA task, including generative question answering, in which an answer is generated that may be a single word or a short phrase, and ranking-based question answering, in which the algorithm is given a predefined list of answers to choose from that are designed to all be plausible.[5] Prior success of VQA technology has been more recently exploited for embodied exploration[7] and navigation of 3-D scenes.[8]

3. Methodology

We establish criteria for determining questions that reflect the normative state of an environment in order to probe for violations of those normative states. These states are, of course, dependent on the particular domain, and in this work, we create them through a manual process by reading source material. To alleviate the burden of manually crafting new criteria for each novel situation, we leave to future work the automatic creation of such normative states by extracting information from source materials.

3.1 Visual Stimuli

We select two sets of visual stimuli for our experiments that vary in their normative states and whether they represent a physical space or a virtual space. The first is a makerspace, a room with tools and equipment (3-D printers, laser cutters, etc.). Safety protocols are critical, and thus detecting anomalies within the space is paramount. A hammer on a bench may be considered misplaced and of lower-level concern, whereas a puddle on the ground may be a hazardous spill. The makerspace is virtually rendered in Unity as part of the Robot Interaction in Virtual Reality (RIVR) platform.[9,10] We manipulated the environment to exhibit a number of potential dangers, then took screenshots from different angles within the environment.

The second visual stimuli depicts an alleyway with atypical properties. This is a real-world location, collected as part of human-robot dialogue experiments found in the Situated Corpus on Understanding Transactions (SCOUT).[11,12] The atypical environment, which includes poor lighting and strange item placements, may evoke natural questions about the normative state and its typical function, thus representative of a space where these elements are unknown prior to exploration.

3.2 Anomaly Categorization

To develop the targeted questions for the VQA, we developed an anomaly categorization for makerspace safety from a lab safety manual by the University of Nevada, Reno Innevation Center.[*] After closely reading the manual, we denoted five broad categories: inventory, equipment, materials, organization, and environment. Table 1 further describes each category and the possible dangers that could result in a violation or critical outlier.

[*]https://www.unr.edu/innevation/makerspace/safety

Table 1 Makerspace anomaly categorization

Inventory	Equipment	Materials	Organization	Environment
- Missing (lost) - Misplaced (found) - Shortage	- Misuse - Malfunction	- Spill - Leak - Broken	- No label - Obstruction - Placement	- Temp. - Noise - Light

These guidelines will serve as a foundation for questioning as well as the development of synthetic scene creation in the virtual makerspace. We note that for this study, we do not create a categorization for the SCOUT environment. This is in part to test the generalization of the makerspace safety to unknown environments.

4. Experiments

4.1 Description Generation Baselines

We first show why using caption description or generic scene description approaches are not adequate for anomaly detection. The caption generation task has seen an explosion of models and accuracy since 2015 with the emergence of neural networks in combination with vision and language tasks[13] (see Wang et al.[14] for a recent survey). One might intuit that an algorithm that can describe an image may be able to therefore describe what is *wrong* in the image. We show, however, that despite the advances in technology, a generic scene description task alone is not adequate for identifying anomalies.

On the makerspace image in Fig. 1 (left), we tested two baselines. The first was the multi-modal framework (MMF) caption generation algorithm,[15] a state-of-the-art suite of natural language processing and computer visions tools released in 2020. The algorithm takes as input a single image, and generates a single natural language sentence describing it. After inputting the makerspace snapshot to MMF, it produced: "a room with a bunch of chairs and umbrellas."

There are no chairs or umbrellas in the room in the image, a clear indicator that MMF could not correctly process the image. Furthermore, MMF's caption generation did not point out any oddities. This may be by design as the caption is meant to describe the state of the image rather than what is *not* in the image; however, this small test suggests that the state-of-the-art caption generation algorithm may not be sufficient for anomaly detection.

We next tested a powerful large pre-trained model (LPTM) called Large Language and Vision Assistant (LLaVA) released in 2023.[16] LPTMs use an image and a natural language prompt as input, and follow the instructions in the prompt to generate an open-ended text response. We created a simple prompt designed to represent a similar instruction as the caption generation task: “Describe the scene.” From the same makerspace image and this prompt, LLaVA generates:

> The scene shows a large, empty room with various items and equipment, such as a table, a chair, and a computer. There are also some barrels and a large container in the room. The room appears to be a laboratory or a workshop, as it contains scientific equipment and tools. The room is empty, which suggests that it may be in the process of being set up or recently cleaned.

The description, while longer than the caption generation response, also did not point out any anomalies in the space. This response may actually be worse than the caption generation’s sparsity because of its contradictory nature. It first delivered an inventory of items in the room, and subsequently declared that the room was empty. This contradiction presents a large risk to an individual seeking this information in order to enter a safe space, where they would be mistakenly led to believe that either the space is empty when it is not, or that the barrel and container are not spilled on the ground along the other missed safety violations.

We selected these two baselines to make the case for targeted questioning for visual anomaly detection. We leave to future work the ability to utilize LLaVA or other LPTMs to implement the same questioning techniques we utilize with the VQA algorithm (Section 4.2), and instead argue that targeted questioning, over a generic request for a caption or description, is necessary.

4.2 VQA Targeted Questioning

We utilize an off-the-shelf VQA algorithm, BLIP (Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation).[17] The BLIP algorithm achieves state-of-the-art results on VQA benchmarks as well as other vision-language tasks, including image-text retrieval and image captioning. BLIP is part of a suite of computer vision and language tools called LAVIS (Library for LAnguage-VISion Intelligence).[18]

We again focused our experimentation on the two images in Fig. 1. Starting with the virtual makerspace, we curated a set of 10 “yes-no” questions based on the image, intending to discover as many anomalies as possible. The questions were designed to probe features of the categorization within the scope of the image. For example, from the Organization→Obstruction category, there should be no obstructions on the ground around items including fire extinguishers, and since there is a fire extinguisher present in the image, we designed the question, “Is anything obstructing the ground around the fire extinguisher?” The questions are listed in Table 2, which also includes the ground truth answer for each visual stimuli and an indicator if the VQA answered the question incorrectly.

Table 2 Ten questions about anomalies. The asterisk (*) indicates the VQA answered incorrectly.

Question	Makerspace answer	SCOUT answer
Is anything obstructing the doorway?	no*	yes*
Is the door left open?	yes	yes
Is anything obstructing the ground around the fire extinguisher?	yes	no*
Is any liquid spilled on the ground?	yes*	no
Is anything inside the caution tape on the ground?	yes	no
Is anything on top of the flammable liquid cabinet?	yes*	no
Are tools left on the table?	yes	no
Are there any items misplaced?	yes	no
Are any items being misused?	no	no
Are any items malfunctioning?	no	no

We implemented a preliminary follow-up routine where, if the VQA answered “yes,” a subsequent question was triggered that specifically requested what exactly in the environment was anomalous. For example, if the VQA answered “yes” to the question, “Is anything obstructing the ground around the fire extinguisher?” The follow-up question triggered was, “What is obstructing the ground around the fire extinguisher?” Due to the wide range of possible answers that could be produced by this open-ended line of questioning, we treat the responses as observations in this work.

The BLIP-VQA algorithm achieved 70% accuracy on these questions in the makerspace, outperforming a random baseline (50% representing chance of the “yes-no" questions). The anomalies overlooked were the liquid spill, the placement of items on top of the flammable liquid cabinet, and that the doorway was not obstructed.

We tested the same 10 questions on the image from SCOUT in Fig. 1 (right) and achieved 80% accuracy. Because these questions were designed for makerspace safety, they are not assessing the anomalies within the SCOUT space, but rather the VQA's ability to answer questions in the atypical environment. This percentage represents precision, rather than recall. While most of the questions were answered correctly, the anomalies *not* asked about that appear in the image were not reflected in the score.

5. Discussion

While we are unable to make broad statements of impact given our small sample size, we discuss several points in our investigation that remain open in ongoing and future work. It is difficult to determine if the errors are due to a computer vision failure at the image level, a language failure at the question understanding level, or a reasoning failure at the algorithmic level. In the case of the question asking if any liquid spilled on the ground, did the computer vision not recognize the oil from the red barrel as a liquid spilling? Did the VQA not understand what a spill is? Or did the algorithm that attempts to reconcile the two fail to do so? The logic of VQA remains "under the hood." In future work, we will stress test and ablate different scenarios to attempt to reveal these differences.

The majority of the questions explicitly name an item of interest for the focus of the analysis (e.g., "doorway") and then a question about it or its surroundings (e.g., "obstruction"). These questions assess the co-location of two objects (in front of, on top of, next to, etc.) and require the computer vision to correctly identify both named items, as well as determine their positional relationship as extracted from the natural language question. All of the incorrect answers in the makerspace stimuli are of this co-location type and may represent shortcomings at the computer vision level (i.e., Can the model identify a "flammable liquid cabinet"?) Does the word "obstruction" generalize such that the model interprets the word to refer to any object? Furthermore, does the fidelity of the simulated environment affect recognition? BLIP is trained on photographic images, although prior work has shown VQA to be effective on animated pictures when trained for such.[5] More recently, work has investigated VQA's performance on synthetically generated environments.[19] There are benefits to training on synthetic environments if we ultimately envision the VQA as assisting in unseen environments. On the other hand, while the SCOUT image is photographic, the lighting, resolution, and camera angle differ from the high-quality photographic images these models are typically trained on. This too may pose a challenge for adapting to unseen environments.

Some of the questions—namely, those dealing with item misplacement, misuse, malfunction—involve implicit knowledge about the item and its typical placement, use, and function. The VQA answered these questions correctly; however, there is no evidence that the VQA has ever been trained on makerspace data, or that it could know typical uses for every item in the entire image. These questions ask for an anomaly detection of a particular type (e.g., anomalies dealing with item misuse.) In subsequent experimentation, more questions will be asked to enumerate possible types of misuse and help focus the VQA analysis.

Finally, we note that some of our questions can have more than one correct answer. To the fire extinguisher obstruction question, for example, there are two such items: a small gas canister and a large oil barrel. Answering in the affirmative is correct, but we envision the follow-up questioning as significant in fully exploring the scene. When dealing with safety protocols, it is not enough to identify the hazard; we must also seek to resolve it. If the VQA line of questioning could respond with the exact cause of the anomaly, recovery actions could be taken by a robot, either by picking up and relocating objects or cleaning them up. In the case of this question, the VQA correctly responds that there is something by the fire extinguisher; however, the follow-up response is "ladder." This returns us to our earlier discussion point about understanding what the VQA is failing under the hood, which we hope to address in experimentation and design of questions from the categorization. We also need to examine whether the categorization and visual stimuli inputs need to be refined.

6. Conclusions and Future Work

In this report, we showed how the use of targeted questioning can more accurately assess anomalies in test environments than image captions or descriptions alone. In particular, this approach demonstrated how manually crafted environments and questions achieved 70% accuracy using an off-the-shelf VQA algorithm, and 80% precision for an out-of-domain image.

We have several research avenues for future work. First, we aim to automate the creation of anomalous categories from source documents. This could be done by applying natural language processing information extraction techniques to documents. As a starting point, we will compare the automatically extracted categories of the makerspace against our manual categories. We will further curate a categorization for the SCOUT environment based on documents that describe the expectations of that environment. Second, we plan to conduct more rigorous question asking over more images, and design images and questions to test the various capabilities (i.e., computer vision, natural language understanding, and VQA.) To substantiate this effort, we will propose new evaluation tasks and metrics for measuring the efficiency of a question-answering interaction between a human and robot based on these questions.

7. References

1. Nagatani K, Kiribayashi S, Okada Y, Otake K, Yoshida K, Tadokoro S, Nishimura T, Yoshida T, Koyanagi E, Fukushima M, Kawatsuma S. Emergency response to the nuclear accident at the Fukushima Daiichi nuclear power plants using mobile rescue robots. Journal of Field Robotics. 2013;30(1):44–63.

2. Kanazawa K, Sato N, Morita Y. Considerations on interaction with manipulator in virtual reality teleoperation interface for rescue robots. IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN); 2023.

3. Jiang X, Xie G, Wang J, Liu Y, Wang C, Zheng F, Jin Y. A survey of visual sensory anomaly detection. arXiv preprint arXiv:2202.07006; 2022.

4. Lukin SM, Sharma R, Bellissimo M. Learning to understand anomalous scenes from human interactions. DEVCOM Army Research Laboratory (US); 2023. Report No.: ARL-TR-9624.

5. Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick CL, Parikh D. VQA: visual question answering. 2015 International Conference on Computer Vision (ICCV); 2015. p. 2425–2433.

6. 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. 2017;163:21–40.

7. Das A, Datta S, Gkioxari G, Lee S, Parikh D, Batra D. Embodied question answering. 2018 Conference on Computer Vision and Pattern Recognition (CVPR); 2018. p. 1–10.

8. Ma X, Yong S, Zheng Z, Li Q, Liang Y, Zhu SC, Huang S. SQA3D: situated question answering in 3D scenes. International Conference on Learning Representations (ICLR); 2022.

9. Higgins P, Kebe GY, Darvish K, Engel D, Ferraro F, Matuszek C. Towards making virtual human-robot interaction a reality. 3rd International Workshop on Virtual, Augmented, and Mixed-Reality for Human-Robot Interactions; 2021.

10. Higgins P, Barron R, Matuszek C. Head pose for object deixis in VR-based human-robot interaction. 31st IEEE International Conference on Robot and Human Interactive Communication; 2022.

11. Lukin SM, Pollard KA, Bonial C, Hudson T, Artstein R, Voss C, Traum D. Navigating to success in multi-modal human-robot collaboration. IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN); 2023.

12. Marge M, Bonial C, Lukin S, Voss C. Bot language (Summary Technical Report, Oct 2016–Sep 2021). DEVCOM Army Research Laboratory (US); 2023. Report No.: ARL-TR-9656.

13. Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: a neural image caption generator. IEEE Conference on Computer Vision and Pattern Recognition; 2015.

14. Wang H, Zhang Y, Yu X. An overview of image caption generation methods. Computational Intelligence and Neuroscience; 2020.

15. Singh A, Goswami V, Natarajan V, Jiang Y, Chen X, Shah M, Rohrbach M, Batra D, Parikh D. MMF: a multimodal framework for vision and language research. https://github.com/facebookresearch/mmf; 2020.

16. Liu H, Li C, Wu Q, Lee YJ. Visual instruction tuning. arXiv:2304.08485; 2023.

17. Li J, Li D, Xiong C, Hoi S. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. 2022 International Conference on Machine Learning (ICML); 2022.

18. Li D, Li J, Le H, Wang G, Savarese S, Hoi SC. LAVIS: a One-stop Library for Language-Vision Intelligence. Association for Computational Linguistics; 2023. p. 31–41.

19. Cascante-Bonilla P, Wu H, Wang L, Feris RS, Ordonez V. SimVQA: exploring simulated environments for visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; p. 5056–5066.

List of Symbols, Abbreviations, and Acronyms

3-D	three-dimensional
BLIP	Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation
LAVIS	A Library for Language-Vision Intelligence
LLaVA	Large Language and Vision Assistant
LPTM	Large Pre-Trained Model
MMF	multi-modal framework
RIVR	Robot Interaction in Virtual Reality
SCOUT	Situated Corpus on Understanding Transactions
VQA	visual question answering

1 (PDF)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA
1 (PDF)	DEVCOM ARL FCDD-RLB-CI TECH LIB
1 (PDF)	DEVCOM ARL FCDD-RLA-IC SM LUKIN
1 (PDF)	UMCP R SHARMA