VQA-Aid: Visual Question Answering for Post-Disaster Damage Assessment and Analysis

Preprint · June 2021

DOI: 10.48550/arXiv.2106.10548

2 authors:

Argho Sarkar
University of Maryland, Baltimore County
9 PUBLICATIONS **446** CITATIONS
SEE PROFILE

Maryam Rahnemoonfar
University of Maryland, Baltimore County
116 PUBLICATIONS **2,040** CITATIONS
SEE PROFILE

VQA-AID: VISUAL QUESTION ANSWERING FOR POST-DISASTER DAMAGE ASSESSMENT AND ANALYSIS

Argho Sarkar, Maryam Rahnemoonfar

Bina Lab, University of Maryland, Baltimore County
Maryland, USA

ABSTRACT

Visual Question Answering system integrated with Unmanned Aerial Vehicle (UAV) has a lot of potentials to advance the post-disaster damage assessment purpose. Providing assistance to affected areas is highly dependent on real-time data assessment and analysis. Scope of the Visual Question Answering is to understand the scene and provide query related answer which certainly faster the recovery process after any disaster. In this work, we address the importance of *visual question answering (VQA)* task for post-disaster damage assessment by presenting our recently developed VQA dataset called *HurMic-VQA* collected during hurricane Michael, and comparing the performances of baseline VQA models.

Index Terms— Visual Question Answering, Post-Disaster Damage Assessment, Hurricane Michael

1. INTRODUCTION

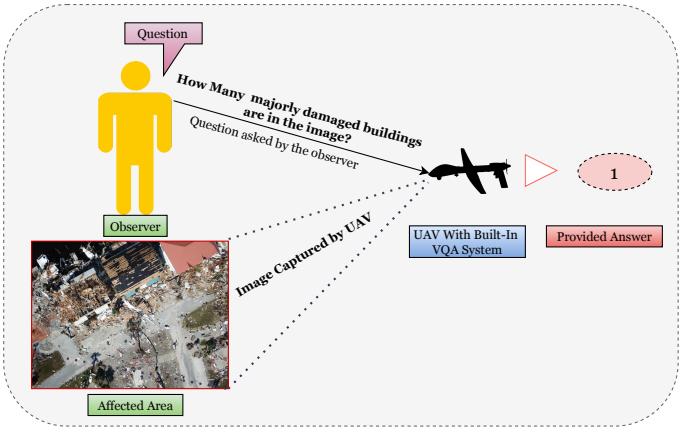

Fig. 1. *VQA-Aid*: At first, a UAV with built-in VQA system captures the images from the affected region and an observer asks a question relevant to the scenario. Finally, after the analysis, the device produces responses.

This work is partially supported by Microsoft.

Visual Question Answering (VQA) is a complicated multimodal research problem in which the aim is to address an image-specified question. Thus, to find the right answer, VQA systems need to model the question and image (visual content). Visual Question Answering is regarded as a cognitive activity that separates it from other perceptual activities, such as the classification of images. For providing answers based on questions in natural language, a VQA model needs to identify the relevant objects from the images, recognize the attributes and find out the interactive relationships among several objects. This high-level scene understanding has the potential to advance the decision support systems for post-disaster damage assessment. Answers from the questions such as "What is the condition of the road?", "How many buildings are damaged?" provides vital information that assists and faster the recovery process which could save many lives. Additionally, the management and the distribution of limited resources can be allocated optimally with the information from the VQA system. However, the success of any VQA model depends on the task-specific data. As the collection of the data is laborious as well as risky due to difficulties to enter the affected areas because of many adverse conditions such as damaged roads, flooded areas, etc., an automated system such as UAV integrated with the VQA module, trained on disaster specific dataset, can be implemented for damage assessment purpose. Understanding the scarcity of VQA datasets for post-disaster damage assessment, we develop a VQA dataset namely *HurMic-VQA* collected after the *Hurricane Michael*. Figure 1 represents the *VQA-Aid* framework in which we showed how the VQA task can be introduced as an assistant tool for disaster assessment that enables us to make the right decision at any time.

Although several datasets are provided for post-disaster damage assessment purposes. Most of those datasets [1, 2, 3] contain satellite images and images collected from social media. However, in [4] authors provide high resolution UAV images. Satellite images are usually captured from high altitudes therefore they have low resolution. Our *HurMic-VQA* dataset contains high resolution UAV images. In most cases, tasks related to the available datasets for natural disaster are limited to classification [5, 6] and semantic segmentation [5, 7, 8]. Visual Question Answering practice has not been considered

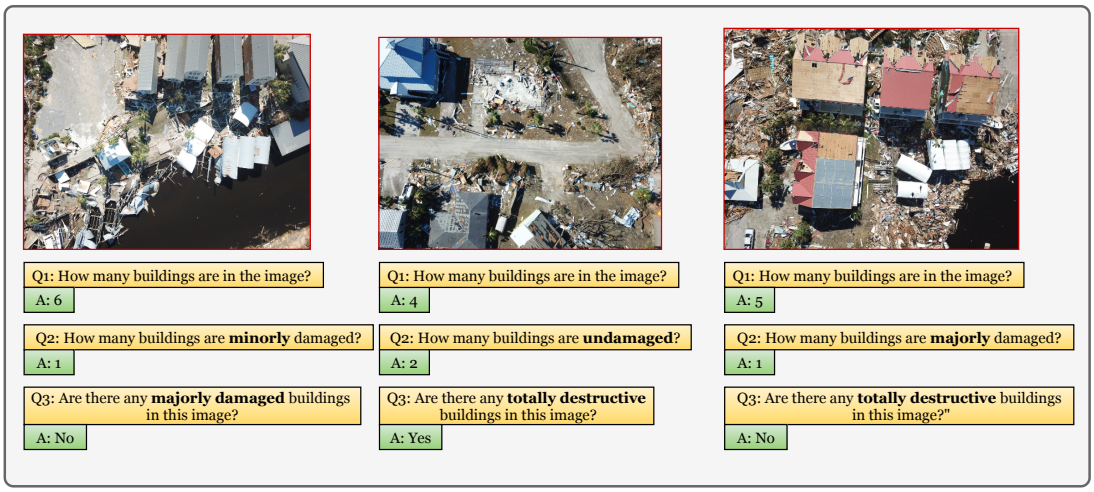

Fig. 2. For all the images, *Q1* represents the *simple counting* question. *Q2* and *Q3* are the reflection of *complex counting* and *yes/no* type of questions. We aim to count the object of a particular attribute in *complex counting* questions (e.g. number of *majorly/minorly* damaged buildings instead of total number of building).

much for the post-disaster damage assessment under climate change issue. To the best of our knowledge, this is the first work addressing VQA tasks based on UAV imagery in climate issues. Substantial research efforts have been made on the development of VQA algorithms [9, 10, 11, 12, 13, 14, 15, 16, 17] in the computer vision and natural language processing communities on many datasets [18, 10]. In these methods, different approaches for the fined-grained fusion between semantic image and question features have been proposed [9, 16, 13, 17]. However, the implementation of VQA algorithms for UAV imagery is complex compare to the other datasets. The representation of UAV images is vertical which is opposite from the everyday images. Differentiating among several objects from a high altitude makes it difficult even for a human. The scenario of the affected areas after a disaster makes it more complicated as there exist many noises such as debris from structural damage compare to the pre-disaster condition. No benchmark results of the well-established VQA algorithms have been provided regarding post-disaster damage assessment based on UAV imagery. To address this issue, we compare the baseline VQA models, in this work, on our dataset. Our work is unique for two reasons. Firstly, we introduce the VQA dataset for post-disaster damage assessment based on UAV imageries, and finally, we conduct a comprehensive study of the performances of baseline VQA algorithms on our dataset.

2. DATASET DESCRIPTION

2.1. Data Collection Process

The dataset is collected with a small UAV platform, DJI Mavic Pro quadcopters, after *Hurricane Michael*. The dataset consists of video and imagery taken from several flights at Ford Bend County in Texas and other directly impacted areas. All the images are high in resolution, i.e., 4000×3000. The damage and debris situation after the hurricane is presented in Figure 2. Though several objects are present, most of the images include debris and buildings. The buildings include both residential and non-residential structures. Table 2 shows the object types with different attributes. While generating the questions. these attributes are considered. In this work, we are interested in investigating the structural damage condition for buildings by asking questions for given images.

2.2. Question Type

Questions are grouped into a three-way category of questions, namely *"Simple Counting"*, *"Complex Counting"*, and *"Yes / No"*. We mainly ask the number of presence of an object in *"Simple Counting"* problem regardless of the associated attribute (e.g. *How many buildings are in the images?*). *Yes / No* type questions concentrate on examining whether an object's particular attribute is present. Finally, *complex counting* type of query is explicitly intended to count the existence of a specific attribute of an object (e.g. *How many **majorly damaged** buildings are in the images?*). A total of 3197 images are available and each image is connected to all of the 3 types of questions. Figure 2 represents these three types of questions.

3. METHOD

The baseline models: *simple baseline* and *Multimodal Factorized Bilinear (MFB) baseline* [16] have been considered for this task and all of these models are configured according to our *HurMic-VQA* dataset. The main pipeline for the

Table 1. Accuracy Results from Baseline VQA Models

Mode of Feature Combination	Loss Function	Data Type	Overall Accuracy	Accuracy for "Simple Counting"	Accuracy for "Complex Counting"	Accuracy for "Yes/No"
Concatenation [9]	Cross Entropy	Training	0.59	0.6	0.56	0.65
		Validation	0.57	0.58	0.54	0.60
		Testing	0.55	0.56	0.53	0.59
	KL Divergence	Training	0.6	0.6	0.56	0.67
		Validation	0.58	0.6	0.54	0.61
		Testing	0.55	0.56	0.53	0.59
Point-wise Multiplication [10]	Cross Entropy	Training	0.59	0.6	0.55	0.64
		Validation	0.58	0.61	0.54	0.61
		Testing	0.57	0.56	0.53	0.65
	KL Divergence	Training	0.6	0.6	0.56	0.67
		Validation	0.59	0.6	0.54	0.66
		Testing	0.58	0.56	0.53	0.68
MFB Module [16]	Cross Entropy	Training	0.59	0.6	0.56	0.64
		Validation	0.58	0.6	0.54	0.63
		Testing	0.56	0.56	0.53	0.63
	KL Divergence	Training	0.59	0.6	0.56	0.63
		Validation	0.57	0.6	0.54	0.60
		Testing	0.57	0.56	0.53	0.64

Table 2. Object with associated Attributes

Object	**Associated Attribute**
Building	Total Destructive, Majorly Damaged, Minorly Damaged, No damage
Road	Covered with Debris, Flooded, Undamaged
Water	Covered with Debris, Flood Water, Clean Water
Pools	Damaged, Undamaged

aforementioned baseline VQA models consists of image feature extraction, semantic representation of question, and fine-grained combination of these two features. For image and question feature extraction, respectively, VGGNet (VGG 16) and Two-Layer LSTM are taken into account. Image feature vector I, $I \in \mathbb{R}^m$ where m represents the dimension of image vector and semantic question feature Q, $Q \in \mathbb{R}^n$ where n represents dimension of question vector, are combined in a *simple baseline* method by both concatenation and point-wise multiplication. 1024-D image feature vector (from last pooling layer) and 1024-D question vector (from the last word of Two-Layer LSTM) are considered for our study.

For the *MFB baseline* approach, authors in [16] proposed the MFB module for a fine-grained combination between image and question feature. The MFB module consists of two phases: expanding and squeezing. Image and question feature vector are multiplied point-wise in the expanding process, followed by a dropout layer. In the squeezing step, sum pooling is considered, followed by power and *L-2* normalization layers.

Fully-connected and softmax layers are taken into account after the fine-grained combination of the two features in all approaches to model the answers. Given a question and an image, the models will predict the answer to the question by formulating the problem as a classification task (for a given set of answers).

4. EXPERIMENT AND RESULT

After the image and question feature extraction from VGGNet and LSTM layer respectively, three modes of feature combination criteria are considered. 1024 dimensional image and question feature vector are combined by concatenation, point-wise multiplication , or MFB module. By considering both cross-entropy and KL divergence loss, all the models are optimized by stochastic gradient descent (SGD) with batch size 16. The dataset is split into three sets namely training, validation and testing with 60%, 20%, 20% ratio respectively. In the training phase, models are validated by validation dataset via *early stopping* criterion with patience 30.

Our motivation for developing an efficient UAV imagery-based VQA model for post-disaster damage assessment comes from the performances of baseline models in Table 1. Accuracy is the performance metric that we consider for the VQA task to compare the baseline models. We consider top-1 accuracy for the comparison purpose. If the ground-truth matches the output (which has the highest probability) from a model, the accuracy for any image is 1, otherwise it is 0. Overall accuracy for all these models lies between .54 and .60 which indicates that models hardly understand the scenario for a given question. *Yes/No* type question has higher accuracy compared to other question types. Furthermore, we highlight the difficulties in dealing with the *complex counting* by comparing the accuracy between *simple counting* and *complex counting* problems. Accuracy for the *complex counting* problem is lower among all the question types. This result highlight the performances of baseline VQA models on our dataset.

5. CONCLUSION

Our aim for this analysis is to study the baseline VQA frameworks by presenting our *HurMic-VQA* dataset for post-

disaster damage assessment. This study also upholds the importance of implementing Visual Question Answering task for post-disaster damage assessment. We only consider a subset of our dataset that targets only one type of object for this work. From the baseline results, we can understand the importance of developing an effective VQA algorithm for post-disaster damage assessment.

6. REFERENCES

[1] Benjamin Bischke, Patrick Helber, Christian Schulze, Venkat Srinivasan, Andreas Dengel, and Damian Borth, "The multimedia satellite task at mediaeval 2017.," in *MediaEval*, 2017.

[2] Bischke Benjamin, Helber Patrick, Zhao Zhengyu, Borth Damian, et al., "The multimedia satellite task at mediaeval 2018: Emergency response for flooding events," 2018.

[3] Ritwik Gupta, Richard Hosfelt, Sandra Sajeev, Nirav Patel, Bryce Goodman, Jigar Doshi, Eric Heim, Howie Choset, and Matthew Gaston, "xbd: A dataset for assessing building damage from satellite imagery," *arXiv preprint arXiv:1911.09296*, 2019.

[4] Maryam Rahnemoonfar, Tashnim Chowdhury, Argho Sarkar, Debvrat Varshney, Masoud Yari, and Robin Murphy, "Floodnet: A high resolution aerial imagery dataset for post flood scene understanding," *arXiv preprint arXiv:2012.02951*, 2020.

[5] Ritwik Gupta, Bryce Goodman, Nirav Patel, Ricky Hosfelt, Sandra Sajeev, Eric Heim, Jigar Doshi, Keane Lucas, Howie Choset, and Matthew Gaston, "Creating xbd: A dataset for assessing building damage from satellite imagery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 10–17.

[6] Christos Kyrkou and Theocharis Theocharides, "Deep-learning-based aerial image classification for emergency response applications using unmanned aerial vehicles.," in *CVPR Workshops*, 2019, pp. 517–525.

[7] Tim GJ Rudner, Marc Rußwurm, Jakub Fil, Ramona Pelich, Benjamin Bischke, Veronika Kopačková, and Piotr Biliński, "Multi3net: segmenting flooded buildings via fusion of multiresolution, multisensor, and multitemporal satellite imagery," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 702–709.

[8] Tashnim Chowdhury, Maryam Rahnemoonfar, Robin Murphy, and Odair Fernandes, "Comprehensive semantic segmentation on high resolution uav imagery for natural disaster damage assessment," in *2020 IEEE International Conference on Big Data (Big Data)*, 2020, pp. 3904–3913.

[9] Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus, "Simple baseline for visual question answering," *arXiv preprint arXiv:1512.02167*, 2015.

[10] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh, "Visual question answering," in *ICCV*, 2015.

[11] Kan Chen, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu, and Ram Nevatia, "Abc-cnn: An attention based convolutional neural network for visual question answering," *arXiv preprint arXiv:1511.05960*, 2015.

[12] Kevin J Shih, Saurabh Singh, and Derek Hoiem, "Where to look: Focus regions for visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4613–4621.

[13] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola, "Stacked attention networks for image question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 21–29.

[14] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *EMNLP*, 2016.

[15] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang, "Hadamard product for low-rank bilinear pooling," *arXiv preprint arXiv:1610.04325*, 2016.

[16] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1821–1830.

[17] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh, "Hierarchical question-image co-attention for visual question answering," in *NeurIPS*, 2016.

[18] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus, "Indoor segmentation and support inference from rgbd images," in *Computer Vision – ECCV 2012*, Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, Eds., Berlin, Heidelberg, 2012, pp. 746–760, Springer Berlin Heidelberg.