Visual Question Answering: A Deep Interactive Framework for Post-Disaster Management and Damage Assessment

Argho Sarkar [1] **Maryam Rahnemoonfar** [1]

Abstract

Each natural disaster has left a trail of destruction and damage, which must be managed very effectively to reduce the disaster's impact. Lack of proper decision in the post-disaster managerial level can increase human suffering and waste a great amount of money. Our objective is to incorporate a deep interactive approach in the decision-making system especially in a rescue mission after any natural disaster to understand the condition of the damages and to take data-driven steps for the systematic distribution of the limited resources and accelerating the recovery process. We believe that Visual Question Answering (VQA) is the finest way to address this issue. In visual question answering, a query-based answer regarding the situation of the affected area can add value to the decision-making system. Our main purpose of this study is to develop a Visual Question Answering model for post-disaster damage assessment purposes. With this aim, we collect the images by UAV (Unmanned Aerial Vehicle) after Hurricane Harvey and develop a dataset that includes the questions that are very important in the decision support system after a natural disaster. In addition, We propose a supervised attention-based approach in the modeling segment. We compare our approach with the two other baseline attention-based VQA algorithms namely Multimodal Factorized Bilinear (MFB) and Stacked Attention Network (SAN). Our approach outperforms in providing answers for several types of queries including simple counting, complex counting compares to the baseline models.

1. Introduction

Disaster management can be defined as accountable organization and management for dealing with all humanitarian aspects, in particular, response and recovery after emergencies to mitigate the impact of disasters. As a result of the significant increase in the earth's surface temperature, global warming has had a dramatic effect in recent years, resulting in various calamities such as hurricanes, catastrophic floods, and wildfires. After these catastrophic events, post-disaster managerial work needs a faster as well as an interactive way to evaluate the impact of the disaster. An in-depth understanding of the situation is very crucial to save many lives and the utmost utilization of the limited resources. Any delay in the decision-making system can enhance human suffering and waste an abundance amount of money. Researchers have shown if we save only 1-minute in the response phase we can save 1000 days in the reconstruction phase. To address this issue and bring the acceleration in the recovery process after any natural disaster, we introduce a visual question answering framework for post-disaster damage assessment. Implementation of the VQA pipeline depends on a large amount of image data. There are various sources from where we can obtain image data after a disaster. The collection of image data from some sources are risky due to the adverse situation during calamities. Unmanned Aerial Vehicle (UAV) is one of the fastest and safest mediums to collect image data. Thus our VQA framework relies on UAV imagery.

In any VQA task, a model needs to detect objects as well as classify their attributes and figure out the interactive relationship to provide answers. This high-level scene understanding has the potential to advance the decision support system for post-disaster damage assessment, especially in the rescue mission. In rescue missions, an interactive approach like visual question answering is highly appreciated for providing query-based answers. In this study, VQA for post-disaster damage assessment provides answers related to queries that occurred due to the flood after any hurricane. "How many buildings are flooded?", "Is the road flooded?", "What is the overall condition of the given image?" are some examples. Answers from those questions certainly provide the information regarding the condition of the affected areas

[*]Equal contribution [1]Department of Information Systems, University of Maryland, Baltimore County, USA. Correspondence to: Argho Sarkar <asarkar2@umbc.edu>, Maryam Rahnemoonfar <maryam@umbc.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

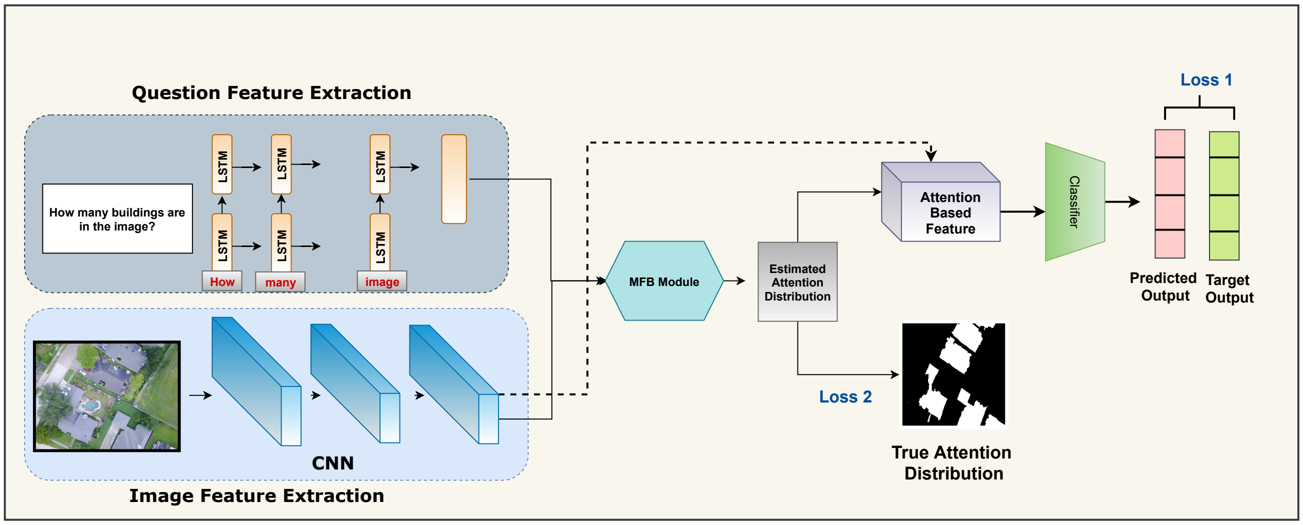

Figure 1. Proposed Supervised Attention-Based Visual Question Answering Model for Post-Disaster Damage Assessment

to the rescuers which assists them to estimate the damages and provide direction to take action. Many other computer vision tasks such as image segmentation (Doshi et al., 2018; Rudner et al., 2019; Zhu et al., 2020), objection detection (Chen et al., 2018; Turker & San, 2004)are considered for disaster management based on aerial (Chowdhury et al., 2020; Fujita et al., 2017; Kyrkou & Theocharides, 2019; Rahnemoonfar et al., 2021; Zhu et al., 2020) and satellite imagery (Chen et al., 2018; Christie et al., 2018; Daudt et al., 2018; Doshi et al., 2018; Gupta et al., 2019; Rudner et al., 2019) . These tasks are not fully effective in terms of providing high-level information and making interacting with rescuers. Though semantic segmentation and objection detection tasks provide a good understanding of the affected areas, they are not suitable for rescue missions due to the lack of interaction. The majority of rescuers in rescue operations are laymen. It might be difficult for certain individuals to interpret the findings and take action according to them. Visual Question Answering (VQA) is an interactive task in which anyone can obtain information in natural language by asking a question in natural language. This is the reason we introduce the task of visual question answering for post-disaster management damage assessment purposes in this study. To do so, we present the VQA dataset and a novel VQA algorithm to address the issue. To the best of our knowledge, this is the first work addressing VQA tasks in this application.

As Visual Question Answering (VQA) is a complicated multimodal research problem in which the aim is to address an image-specified question, it needs to model the question and image (visual content). However, developing UAV imagery-based VQA algorithm for the post-disaster damage assessment purpose is very challenging compare to other datasets. Representation of UAV images refers to vertical representation which is different compare to the horizontal representation captured by traditional digital cameras. Top-view pictorial representations from UAV make it very difficult to distinguish between several objects even for a human as the objects of interest become relatively small. In the case of damage assessment, this degree of complexity gets much higher due to noises come from many sources such as structural debris after any natural disaster. Therefore, special care needs in the modeling part to achieve success in providing answers from the UAV-based VQA system. The success of a VQA system for post-disaster damage assessment depends on how well it can identify the relevant parts of an image based on a question and provide an answer. Attention-based VQA systems (Anderson et al., 2018; Fukui et al., 2016; Lu et al., 2016; Yang et al., 2016; Yu et al., 2017) have shown remarkable performance over many VQA datasets (Antol et al., 2015; Silberman et al., 2012). In most attention-based VQA models (Yang et al., 2016; Yu et al., 2017), attention over the images learns only by minimizing the loss between the actual and predicted answer. Due to the complexity of the UAV images, in our case, the estimated distribution of attention weights over images could not give importance to the relevant part which leads to an incorrect answer. To improve the attention, we present a supervision technique where a true distribution of attention map surrogate the supervision for obtaining the more relevant attention weight distribution in a multitask learning manner. Figure 1 represents our VQA model, where we provide true attention distribution by masking the irrelevant portions of the image for a given question. Learning this extra-label can teach the model where to look into the image based on questions for providing the answer. In our multitasking approach, we minimize the categorical loss (i.e., loss between the actual and predicted answer) along with loss between estimated and true attention weight. We assume that providing supervision for achieving more relevant attention increases the performance of VQA systems in complex scenarios like ours.

2. Dataset for VQA

2.1. Data Collection

The data collection process had taken place after the *Hurricane Harvey*. *Hurricane Harvey* was a Category 4 hurricane that hit Texas and Louisiana in August 2017. We take the advantage of an unmanned aerial vehicle (UAV) platform, DJI Mavic Pro quadcopters, to capture images and videos from the affected areas. The data were collected from Ford Bend County in Texas and other directly impacted areas between August 30 - September 04, 2017, from several flights. All our images are high in resolution, 4000×3000, which makes them unique from other natural disaster datasets.

2.2. Types of Question

The selection of question type is very important so that it can justify the purpose of incorporating the VQA system in a rescue mission. Each type of question should contain information that can provide the rescuers a better understanding of the affected areas so that the distribution of limited resources could be utilized properly and recovery process could be faster.

- At first we need to identify the flooded and non-flooded area. To address this, we include a question that provides an answer related to this query. “What is the overall condition of the entire image?” is an example of this type of question.
- Rescuers need to identify the condition of the road to save lives and operate their mission. To investigate whether the impacted area is reachable by road, rescuers need to understand the road condition before start their move to that specific area. Responses from the questions such as “What is the condition of the road?”, “Is the road flooded?” will serve this purpose.
- Counting the structural entities will provide the intuition of the level of human life risk and magnitude of damages. For instance, if numerous buildings are damaged by flood, the condition of living people in those buildings will be in danger. We separate these counting-related questions into two categories namely simple counting and complex counting. In the *Simple Counting* problem, we ask about an object’s frequency of presence (mainly about building) in an image regardless of the attribute (e.g. *How many buildings are in the images?*).*Complex Counting* is specifically intended to count the number of a particular building attribute (e.g. *How many **flooded / non-flooded** buildings are in the images?*) We are interested in counting only the flooded or non-flooded buildings from this type of query. In comparison to simple counting, a high-level understanding of the scene is important for answering this type of question.

2.3. Types of Answer

Table 1. Possible Answers for Three Types of Question

Question Type	Possible Answer
Simple Counting	{1,2,3,4...}
Complex Counting	{1,2,3,4...}
Condition of Road (sub-category of Condition Recognition)	Flooded , Non-Flooded, Flooded & Non-Flooded
Condition of Entire Image (sub-category of Condition Recognition)	Flooded , Non-Flooded
Yes/No-Type Question (sub-category of Condition Recognition)	Yes, No

Table 1 refers to the possible answers for the three types questions. Most frequent answers for counting problem, in general, are ‘4, 3, 2, 1’ whereas ‘27, 30, 41, 40’ are the less frequent answers. For *Condition Recognition* problem, ‘non-flooded, yes’ are the most common answers.

3. Supervised Attention-Based VQA Model

Due to challenges involved in UAV imagery-based VQA approach, described in section 1, we assume that estimated attention that obtains by only minimizing the error between predicted and ground-truth answers in a classification manner fails to give importance to the most relevant portions over images. Therefore, we provide true attention distribution in the training process so that attention weights can be learned by minimizing the distance between estimated and true attention distribution. We believe that by learning this additional-label (defines as attention loss), we can enhance the performance of VQA algorithms for post-disaster management. In our approach, inference can be made without providing the true attention map.

3.1. True Attention MAP

To provide the true attention distribution, we mask the irrelevant portions from images based on the question. To mask images, we take the advantage of semantic segmentation. All the images of the dataset are annotated pixel-wise with nine classes which include building-flooded, building-non-flooded, road-flooded, road-non-flooded, water, tree, vehicle, pool, and grass. A building is classified as flooded when at least one side of a building touches the floodwater. “Water” class has been considered for representing any natural water body like river and lake. In addition, Road is classified as flooded if the road is submerged under floodwater. If more than 30% area of an image is occupied by

Table 2. Compare the Accuracy (top-1) Between Proposed and Baseline Methods

				Counting problem		Condition Recognition		
VQA Method	Attention Loss	Data Type	Overall Accuracy	Accuracy for 'Simple Counting'	Accuracy for 'Complex Counting'	Accuracy for 'Yes/No'	Accuracy for 'Road Condition'	Accuracy for 'Entire Image Condition'
MFB with Attention (Yu et al., 2017)	X	Validation	0.72	0.32	0.28	0.97	0.97	0.96
	X	Test	0.71	0.27	0.25	0.96	0.95	0.96
SAN (Yang et al., 2016)	X	Validation	0.65	0.29	0.28	0.56	0.97	0.96
	X	Test	0.65	0.32	0.29	0.56	0.95	0.95
With supervised attention(Ours)	MAE	Validation	**0.72**	**0.34**	**0.3**	0.92	**0.97**	**0.98**
		Test	**0.72**	**0.36**	**0.31**	0.8	**0.98**	**0.97**

flood water then that image is classified as flooded, otherwise non-flooded. From semantic segmented images, we then mask the irrelevant portions of images by replacing the pixel value with [0,0,0] considering the RGB channel and highlight the relevant portions by replacing the pixels values with [1,1,1]. For instance, to answer the question like "What is the condition of the road?", the attention should focus on the features related to the road. To do so, we mask the image except for the road portion and consider this as true attention in the training process. In Figure 2, we demonstrate the image-question pairs with associated true attention distribution.

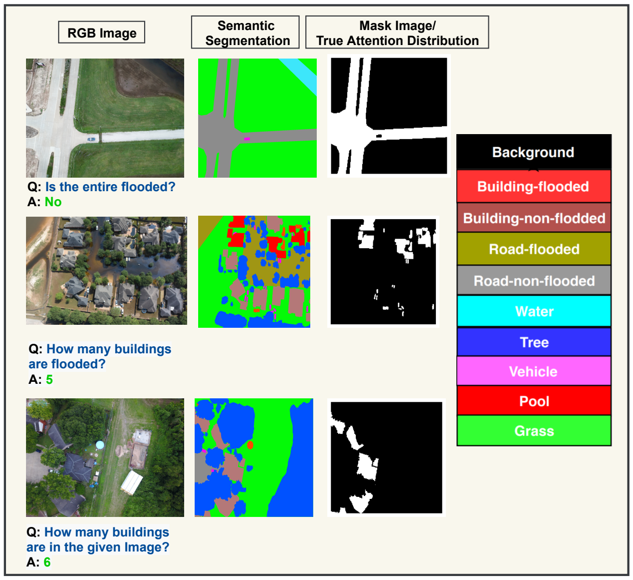
RGB Image
Semantic Segmentation
Mask Image/ True Attention Distribution
Q: Is the entire flooded?
A: No
Q: How many buildings are flooded?
A: 5
Q: How many buildings are in the given Image?
A: 6
Background
Building-flooded
Building-non-flodded
Road-flooded
Road-non-flooded
Water
Tree
Vehicle
Pool
Grass

Figure 2. Overview of the dataset. Each image is associated with a semantic segmented image and each masked image is generated based on the question. Each masked image provides true attention distribution from where a model can learn where it should look into.

3.2. VQA Framework

We first obtain the image feature matrix, I, from the last pooling layer of VGGNet architecture. We resize the RGB images to be 224×224 pixels and extract image feature matrix of size $14 \times 14 \times 512$, where 14×14 is the number of grids/regions in the image and 512 is the dimension of the feature vector for each region. In another way, each grid represents the 16×16 pixel region of the input image. Two-Layer LSTM (Long Short-Term Memory) has been taken into account for textual feature representation. $1024 - D$ semantic question feature vector, Q, from the last LSTM cell of Two-Layer LSTM is considered for our study. Both features are then fused by MFB module described in (Yu et al., 2017). To derive attention from the images, the softmax function is considered. After the attention layer, two branches are considered for the multitasking approach. One branch will try to minimize the distance between true and estimated attention weight matrix by minimizing *loss 2* and the second branch will try to minimize *loss 1* between the true and predicted answer from the VQA classifier. We consider categorical cross-entropy loss for *loss 1*. On the other hand, mean absolute error (MAE) is considered to minimize the distance between true and estimated attention distribution.

4. Results

Table 2 shows the comparison of the performance between the baseline methods and our proposed method. We consider top-1 accuracy for comparison. We can identify that our proposed method performs better in all categories except the 'yes/no' type of question. The increment of the accuracy for simple counting is 9%, 4% compare to MFB and SAN models respectively. Accuracy improves by 6%, 2% for complex counting over MFB and SAN models respectively. Our model shows significant improvement, 24%, for the 'yes/no' type of questions over the SAN model. Performance of supervised attention-based VQA model for providing answers regarding road condition and entire image condition is 3% and 1% more accurate respectively compare to the baseline models.

5. Conclusion

In this study, we present the idea of visual question answering for post-disaster management and damage assessment purposes. We mainly try to establish the importance of the VQA task in a rescue mission after any natural disaster. A supervised attention-based visual question answering algorithm for post-disaster damage assessment based on UAV imagery is presented. Our model shows impressive improvement over the baseline models.

References

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., and Parikh, D. Visual question answering. In *ICCV*, 2015.

Chen, S. A., Escay, A., Haberland, C., Schneider, T., Staneva, V., and Choe, Y. Benchmark dataset for automatic damaged building detection from post-hurricane remotely sensed imagery. *arXiv preprint arXiv:1812.05581*, 2018.

Chowdhury, T., Rahnemoonfar, M., Murphy, R., and Fernandes, O. Comprehensive semantic segmentation on high resolution uav imagery for natural disaster damage assessment. In *2020 IEEE International Conference on Big Data (Big Data)*, pp. 3904–3913, 2020. doi: 10.1109/BigData50022.2020.9377916.

Christie, G., Fendley, N., Wilson, J., and Mukherjee, R. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6172–6180, 2018.

Daudt, R. C., Le Saux, B., Boulch, A., and Gousseau, Y. Urban change detection for multispectral earth observation using convolutional neural networks. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pp. 2115–2118. IEEE, 2018.

Doshi, J., Basu, S., and Pang, G. From satellite imagery to disaster insights. *arXiv preprint arXiv:1812.07033*, 2018.

Fujita, A., Sakurada, K., Imaizumi, T., Ito, R., Hikosaka, S., and Nakamura, R. Damage detection from aerial images via convolutional neural networks. In *2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA)*, pp. 5–8. IEEE, 2017.

Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., and Rohrbach, M. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*, 2016.

Gupta, R., Goodman, B., Patel, N., Hosfelt, R., Sajeev, S., Heim, E., Doshi, J., Lucas, K., Choset, H., and Gaston, M. Creating xbd: A dataset for assessing building damage from satellite imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 10–17, 2019.

Kyrkou, C. and Theocharides, T. Deep-learning-based aerial image classification for emergency response applications using unmanned aerial vehicles. In *CVPR Workshops*, pp. 517–525, 2019.

Lu, J., Yang, J., Batra, D., and Parikh, D. Hierarchical question-image co-attention for visual question answering. In *NeurIPS*, 2016.

Rahnemoonfar, M., Chowdhury, T., Sarkar, A., Varshney, D., Yari, M., and Murphy, R. R. Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access*, 9:89644–89654, 2021. doi: 10.1109/ACCESS.2021.3090981.

Rudner, T. G., Rußwurm, M., Fil, J., Pelich, R., Bischke, B., Kopačková, V., and Biliński, P. Multi3net: segmenting flooded buildings via fusion of multiresolution, multisensor, and multitemporal satellite imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 702–709, 2019.

Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. Indoor segmentation and support inference from rgbd images. In Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., and Schmid, C. (eds.), *Computer Vision – ECCV 2012*, pp. 746–760, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-33715-4.

Turker, M. and San, B. T. Detection of collapsed buildings caused by the 1999 izmit, turkey earthquake through digital analysis of post-event aerial photographs. *International Journal of Remote Sensing*, 25(21):4701–4714, 2004. doi: 10.1080/01431160410001709976. URL https://doi.org/10.1080/01431160410001709976.

Yang, Z., He, X., Gao, J., Deng, L., and Smola, A. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 21–29, 2016.

Yu, Z., Yu, J., Fan, J., and Tao, D. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 1821–1830, 2017.

Zhu, X., Liang, J., and Hauptmann, A. Msnet: A multilevel instance segmentation network for natural disaster damage assessment in aerial videos. *arXiv preprint arXiv:2006.16479*, 2020.