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

Argho Sarkar, Maryam Rahnemoonfar

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

July 10, 2021

Motivation

- Post-disaster management is crucial to mitigate human sufferings after any disaster.
- Any delay in decision making in the post-disaster managerial level can increase human suffering and waste a great amount of money.
- Most of deep learning methods for the post-disaster damage assessment purposes is less interactive and time consuming.

Objective

- 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 impact of disaster and to take necessary steps for the systematic distribution of the limited resources and accelerating the recovery process.
- **Visual Question Answering (VQA)** is the finest way to address this issue.
- Our main purpose of this study is to develop a supervised attention-based Visual Question Answering (VQA) model for post-disaster damage assessment purposes based on UAV imagery.

Dataset Preparation

- **UAV Imagery Collection Process**

- 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.
- All our images are high in resolution, 4000×3000, which makes them unique from other natural disaster datasets.

Dataset Preparation

- **Question Generation Process**

- The selection of question type is very important so that it can justify the purpose of incorporating the VQA system in a rescue mission.
- The selection of question type is very important so that it can justify the purpose of incorporating the VQA system in a rescue mission.

Type of Questions

- **Simple Counting** ask about an object's frequency of presence in an image regardless of the attribute. For example, "How many buildings are in the images?"

- **Complex Counting** is specifically intended to count the number of a particular attribute of an object. For example, "How many **flooded / non-flooded** buildings are in the images?"

- **Condition Recognition** is divided into three sub-categories:
  - *Road Condition* related questions investigate whether the impacted area is reachable by road. "What is the condition of the road?" is an example of this type of questions.
  - *Yes/No* type of questions seek answers between yes or no. For example, "Is the road flooded?"
  - *Entire Image Condition* related questions identify the overall condition. For instance, "What is the overall condition of the entire image?"

Challenges in UAV imagery

- 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.

Supervised Attention-Based VQA Model

- Due to challenges involved in UAV imagery-based VQA approach, 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.
- 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.
- 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.

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.
- 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].

Dataset

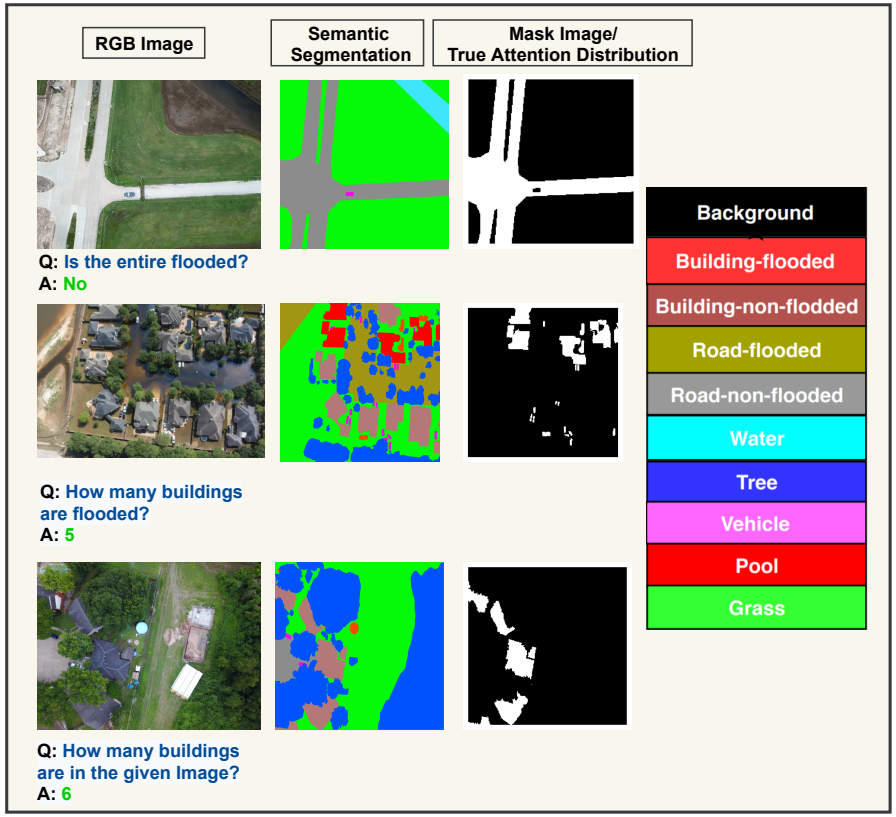

Figure: Overview of the dataset. Each image is associated with a semantic segmented image and each masked image is generated based on the question. Each masked image provides true attention distribution from where a model can learn where it should look into.

Architecture

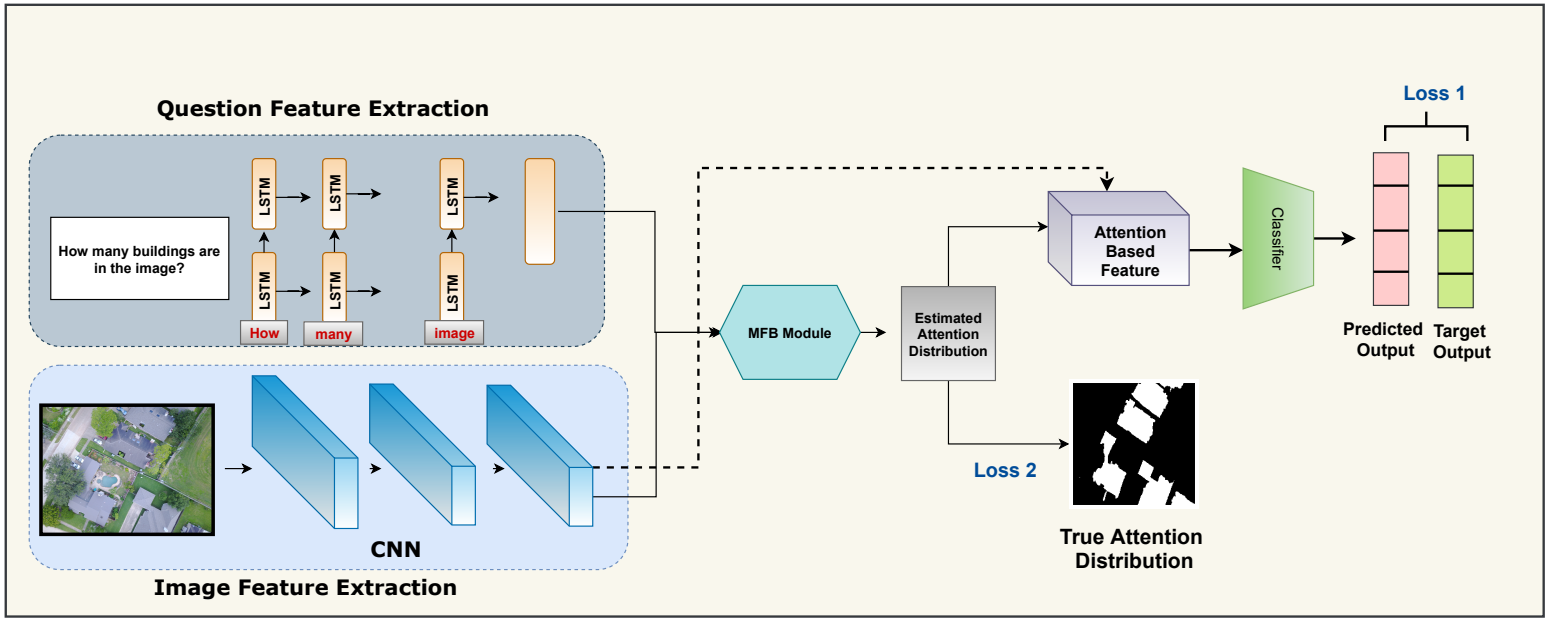

Figure: Figure 2 represents our VQA model, where we provide true attention distribution by masking the irrelevant portions of the image for a given question. Learning this extra-label can teach the model where to look into the image based on questions for providing the answer.

Experimental Results

VQA Method	Attention Loss	Data Type	Overall Accuracy	Counting problem		Condition Recognition		
				Accuracy for 'Simple Counting'	Accuracy for 'Complex Counting'	Accuracy for 'Yes/No'	Accuracy for 'Road Condition'	Accuracy for 'Entire Image Condition'
MFB with Attention	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	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**

- 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.

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.