# A Novel Disaster Image Data-set and Characteristics Analysis using Attention Model

Fahim Faisal Niloy, Arif, Abu Bakar Siddik Nayem, Anis Sarker, Ovi Paul
M. Ashraful Amin, Amin Ahsan Ali, Moinul Islam Zaber†, AKM Mahbubur Rahman
AGenCy Lab, CSE, Independent University, Bangladesh; †CSE, Dhaka University, Bangladesh
niloy9542@gmail.com, [1611041, 1510190, 1521745, 1531144, aminmdashraful, aminali]@iub.edu.bd,
†zabermi@gmail.com, akmmrahman@iub.edu.bd

***Abstract*—The advancement of deep learning technology has enabled us to develop systems that outperform any other classification technique. However, success of any empirical system depends on the quality and diversity of the data available to train the proposed system. In this research, we have carefully accumulated a relatively challenging dataset that contains images collected from various sources for three different disasters: fire, water and land. Besides this, we have also collected images for various damaged infrastructure due to natural or man made calamities and damaged human due to war or accidents. We have also accumulated image data for a class named non-damage that contains images with no such disaster or sign of damage in them. There are 13,720 manually annotated images in this dataset, each image is annotated by three individuals. We are also providing discriminating image class information annotated manually with bounding box for a set of 200 test images. Images are collected from different news portals, social media, and standard datasets made available by other researchers. A three layer attention model (TLAM) is trained and average five fold validation accuracy of 95.88% is achieved. Moreover, on the 200 unseen test images this accuracy is 96.48%. We also generate and compare attention maps for these test images to determine the characteristics of the trained attention model.**

**Keywords**: Disaster Image, Attention Model, Class Activation Map, Three Layer Attention Module.

## I. Introduction

In recent days, natural disasters, i.e., floods, cyclones, droughts, and earthquakes are becoming more common due to climate change, world-wide temperature rise, and pollution. Moreover, population density and socio-economic environments cause human-made disasters that include fire, building collapse, infrastructural damage, road accident, and armed war etc. Situations are getting worse in the developing countries that have very high density populations along with weak socio-economic structures. Generally, thousands are affected by these disasters. Therefore, it is crucial in times of crisis that emergency response workers reach at the affected premises promptly to save human lives and prevent loss. It would be great to have a system that would raise an alert and quantify the degree of damage of any disaster and inform the appropriate authorities based on an automated analysis of the images that are almost available in real-time on various social media. However, the state-of-the art deep learning techniques are not able to classify disaster types from images due to lack of standard disaster datasets. Existing disaster datasets have many limitations such as insufficient categories, imbalanced classes, wrong annotations etc. Therefore, in this paper, we propose an elaborated and standard dataset that has disaster images collected from google, twitter, facebook and other social media sites, online news portals and other standard datasets. Moreover, the proposed dataset also contains images of recent disasters: wild fires in Australia, flood in India, forest fire in Amazon, and many more. We have performed a number of experiments to show that the dataset can help build effective classifier models. Examples of different disaster images are shown in figure 1.


Fig. 1: (a) Fire Disaster, (b) Water Disaster, (c) Infrastructure Damage, (d) Human Damage, (e) Land Disaster and (f) Non-Damage

A number of research has been conducted into classifying different kinds of disasters, both from deep learning and image processing domains in recent years. Different datasets have also been proposed for making learning process effective. Arif et al. [1] have collected and experimented with South Asian Disaster (SAD) images that include disaster images from Bangladesh and other south asian countries. The authors here have observed that the appearance of disaster images of south asia differ in various ways from western disaster images. However, the main limitation of SAD dataset is that the images per class are too limited to be used to train deep learning models.

In the paper [4], the authors extracted a total of 68457

TABLE I: Comparison of Several Existing Datasets

| Authors | Dataset | Number of Classes | Names of the classes | Number of Images | Comments |
|---|---|---|---|---|---|
| Rizk et al. [2] | Home-grown + Sun dataset | 2 | Infrastructure and Natural Disaster | 2344 | Contains only two classes |
| Giannakeris et al. [3] | 3F-emergency dataset | 2 | Fire and Flood Disaster | 12000 | Contains only two classes, Low diversity |
| Muhammad et al. [4] | Ko et al. [5] + Verstock et al. [6] + Chino et al. [7] + Foggia et al. [8] | 2 | Fire Disaster and Non-Damage | 68457 | Contains only two classes |
| Alam et al. [9] | Image4act | 4 | Nepal Earthquake, Ecuador Earthquake, Typhoon Ruby, and Hurricane Matthew | 34562 | Natural disasters only, Limited to narrow geographical regions |
| Arif et al. [1] | South Asia dataset (SAD) | 6 | Fire Disaster, Flood Disaster, Infrastructure, Nature Disaster, Human Damage and Non Damage | 493 | The images per class are very few |
| Mouzannar et al. [10] | UCI dataset | 6 | Fire Disaster, Flood Disaster, Infrastructure, Nature Disaster, Human Damage and Non Damage | 5880 | Non-damage class contains irrelevant images, Small number of images per class, Low diversity, Dataset-bias |
| Niloy et al. | Proposed dataset [11] | 6 | Fire Disaster, Flood Disaster, Infrastructure, Land Disaster, Human Damage and Non Damage | 13720 | Several subcategories, High diversity, Covers broad geographical regions, Natural and man-made disasters, Reduced bias |

images from dataset [6] and Chino dataset [7]. From Foggia dataset [8], they collected video frames. They used an architecture similar to Alexnet [12] and achieved state-of-the-art results.

In the work [3], the authors classified, localized, and estimated severity of different disasters. They used MediaEval [13] dataset for classification and Bow fire dataset [7] for localization. They developed their own dataset named 3F emergency dataset that consisted of flood and fire pictures taken from flicker. Their classification algorithm surpassed other participants in accuracy metric of MediaEval challenge. However, their datasets suffer from inadequate disaster events. They have performed experiments with only flood and fire classes. Rizk et al. [2] proposed a multi-modal two-stage framework that relies on computationally inexpensive visual and semantic features to analyze Twitter data. In this paper, two datasets were used: Home-grown and Sun dataset. Home-grown dataset comprises of Twitter images which only covers damaged infrastructure and natural disaster. Sun dataset was made from several search engines and it also contains infrastructure and natural disaster. So, these datasets contain two categories only. Moreover, the size of their dataset is very small: only 2344 images.

In [13], the researchers present the algorithms that the team deployed to tackle disaster recognition tasks. They made two flood disaster dataset: one of them was made from social media image and the other one from satellite images. GoogleNet architecture was used to train on the images. A major limitation of their dataset is that it only contains flood disaster category. Alam et al. [9] proposed an image filtering module that employs deep neural networks and perceptual hashing techniques to determine whether a newly-arrived image is relevant for a given disaster response context. To train the relevancy filter, 3,518 images were randomly selected from the severe and mild categories. The authors have collected four types of natural disasters: Nepal Earthquake, Ecuador Earthquake, Typhoon Ruby, and Hurricane Matthew. No other regional disaster images are present in the dataset.

Mouzannar et al. [10] proposed a multimodal deep learning framework to identify damage related information from social media posts. This framework combines multiple pretrained unimodal convolutional neural networks that extract features from raw texts and images separately. The framework was evaluated on a homegrown labeled dataset that contains images collected from social media posts. Their dataset (UCI dataset) contains six categories: fire, flood, natural disaster, infrastructure damage, and non-disaster. Though this dataset contains images for various disaster events, their non-disaster class contains lots of irrelevant images, e.g, images of foods, products, jewelries etc. Keeping these images into non-disaster class might result in good overall classification performance but learned models would not distinguish between damaged and undamaged infrastructures. Another limitation is, the deep models trained on UCI dataset do not show well attention localization capability, because the dataset is not much diverse. For example, we observed that the models trained on UCI dataset use fire-trucks in the disaster image as a proxy to classify fire disaster event, which is not expected. Some examples of misplaced attentions are shown in figure 2.

In the paper [14], images posted on social media platforms during natural disasters are analyzed to determine the severity of damage caused by the disasters. The authors collected images of different disasters: Typhoon Ruby, Hurricane Matthew, Nepal Earthquake from internet. The authors also used google search to collect images like damaged building, damaged bridge, damaged road etc. The limitation of the dataset is that it only contains damaged infrastructure images.

Summary statistics of the datasets are shown in table I. From the datasets mentioned above, it is easy to notice that a benchmark dataset for disaster classification is yet to be

published. In most of the literature, a scarcity of benchmark dataset is clearly seen. In summary, the limitations are:

- The datasets do not cover broad regions. A comprehensive dataset should have disaster images from most major regions of the world.
- After training deep learning architectures on several existing disaster datasets, we observe that the classifiers show poor attention localization capability, because the datasets are not diverse enough. That is why the classification accuracy deteriorates.
- Most of the datasets do not contain images having enough challenging scenarios. As a result, the algorithms are prone to misclassification when exposed to semantically similar images. For example, it is common for architectures trained on fire disaster images to misclassify images with high brightness and reddish hue to be a fire disaster event.
- A diverse dataset having good volume of disaster images with wide number of categories and subcategories is yet absent.

To overcome the above limitations, we propose a novel dataset where we have collected images for a number of disaster events that include both natural and non-natural(man-made) disasters from different geographical regions. Also we have carefully hand-picked and annotated several test images which are used in separate attention models to show the efficacy of our proposed dataset. Most specifically our contributions are:

- A novel disaster dataset with 6 disaster categories and 10 subcategories, consisting of a total 13720 images. The detailed statistics of our proposed dataset is shown in table II
- Bounding box annotated images for 200 test images. These are used in attention verification to show improved attention localization capability of classifiers trained on our dataset.
- Detailed characteristics analysis using attention models. We use CAM[15], TLAM[16] to show deep learning classifiers trained with our dataset yield better results compared to existing datasets.

## II. Proposed Dataset Description

A well described and diversified dataset is needed for deep learning systems to perform well in classification. Therefore the objective of this paper is to provide a well defined and diversified dataset for disaster classification. As most of the existing datasets do not contain disaster images from major regions, our purpose is to create a dataset which contains images from all major regions (i.e. western, asian, tropical regions etc.). Existing datasets do not have well organized sub categories. Therefore, another of our objective is to create a dataset that contains well organized sub categories.

Moreoever, a good reason of deep learning networks' poor performance in classifying most existing disaster datasets is that the networks do not focus their attention on disaster related items (e.g. smoke, water etc.). Generally, disaster images have many subjects involved. For example, a cat image may have a cat which is the subject and simple background. A dog image may also have the same scenario. So it is easier to differentiate between a dog and a cat. However, disaster images may have wide range of semantics involved, i.e. there can be images with damaged buildings, crowd of humans, cluttered backgrounds etc. Consequently, it becomes tough for deep learning networks to focus on the correct region of interest. While training with UCI images, we have observed that the classifier learns the fire trucks as features for fire disaster since large number of fire images contain fire trucks. These types of misplaced attention might produce inappropriate features that would result in poor classification performance. For this reason, we have carefully collected thousands of images having wide range of varieties so that the network is forced to learn to focus it's attention on disaster related items. This also helps in improving classification performance which is shown in the experiment section. Paying attention to the appropriate regions also confirms the quality of classifier models.

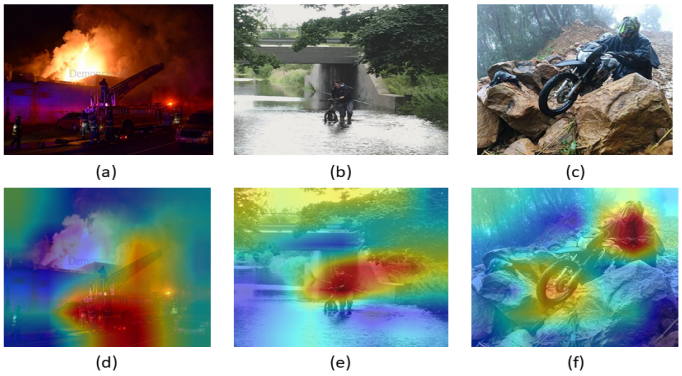



Fig. 2: First row contains input images from different disaster classes, second row contains corresponding attention heatmaps; (a) Fire Disaster (b) Water Disaster (c) Infrastructure Damage (d) Misplaced attention (vehicle); attention should be focused on fire region (e) Misplaced attention (vehicle and human); attention should be focused on water region (f) Misplaced attention (human, motorcycle); attention should be provided to the damaged infrastructure region

We have collected images for three types of disasters: Fire Disaster, Water Disaster, Land Disaster. Additionally, there are two damage related classes: Damaged Infrastructure, Human Damage. The sixth class is Non-Damage where normal images with various infrastructure, natural scene, forest, beach are grouped together. We have also added several sub categories: Urban fire and Wild fire in Fire Disaster category, Landslide and Drought in Land Disaster category etc. Figure 3 shows example images from fire subcategory. Moreover, we have created four subcategories for Non Damage class: Human, Building and Street, Wildlife Forest and Sea. The Non-Damage images are limited to four categories because we

wanted to put a 'negative set' for each disaster category, e.g, the non-damage human sub-category can be considered as a negative set for "Human Damage" category. The same way, non-damage buildings and streets sub category is negative for "Damaged Infrastructure" or "Urban Fire"; non-damage sea is negative for "Water disaster" and non-damage forest can be considered a negative set for both "Land Disaster" and "Wild fire". This is also the reason why the Non-Damage category contains the most number of images. The proposed dataset is made publicly available here [11].

TABLE II: Proposed Dataset Summary

| Category | Sub-Category | Train | Test | Total |
|---|---|---|---|---|
| Damaged Infrastructure | Infrastructure | 1418 | 34 | 1488 |
| | Earthquake | 36 | | |
| Fire Disaster | Urban Fire | 419 | 33 | 966 |
| | Wild Fire | 514 | | |
| Human Damage | | 240 | 32 | 272 |
| Water Disaster | | 1035 | 33 | 1068 |
| Land Disaster | Land Slide | 420 | 33 | 654 |
| | Drought | 201 | | |
| Non Damage | Human | 120 | 35 | 9272 |
| | Building and Street | 4572 | | |
| | Wildlife Forest | 2271 | | |
| | Sea | 2274 | | |
| Total Images | | 13520 | 200 | 13720 |

### A. Disaster Image Collection

We have collected disaster images from different number of sources. Normal buildings, street, forest, and sea images are collected from google and have been put in non disaster category. A large number of different disaster images such as fire, earthquake, tsunami, landslide and flood have been collected from google and popular social media sites such as facebook, twitter etc. We have focused on notable recent disasters like Kerala floods from South India, Japan's tsunami for water disaster, Australian bushfires, California wildfires, Brazil Amazon rain forest wildfires, Hong Kong protests police violence, and so on for fire disaster. Moreover, we have collected a number of disaster images by scrapping different news portals. Some notable examples are: California wildfires [17], Brazil wildfires [18] and many more. In Social media platforms, *hashtag* categorizes similar type of posts or images. We have used *hashtags* to collect image data for different kind of disasters from facebook and twitter. To collect images for human damage, we have focused on Syrian civil war, Yemeni civil war etc. Additionally, we have gathered some of the damaged infrastructure images from news portals [19], facebook, twitter etc. We have taken several building fire, forest fire, damaged infrastructure, tsunami etc. images from SAD [1]. Also we have collected a lot of non disaster images from [20].

After collecting images with above mentioned process, we gathered almost 16000 images in total. However, we had to discard few number of images which were very low resolution, had embedded texts etc. A wide number of images were later discarded in the course of annotation. We finally got a total of 13720 images. The image shapes of our dataset are diverse. During training, the images were resized to $224 \times 224$. The class statistics of the collected images are shown in table II.

(a) Urban Fire (b) Wild Fire

Fig. 3: Subcategories of Fire Disaster

### B. Annotation of Disaster Image Categories

After collecting and cleaning up, the category for each image is determined by our well trained annotators. For this purpose, three human annotators have been trained beforehand on the image classification ideas and methodologies. They learned about different disaster classes and sub-classes. Also, they have gone through different video and image sources to understand the impact of different disasters. Each image from the collected set has been annotated by three annotators separately without any knowledge about the annotation of others.

For Water Disaster category, the defining characteristic is excessive amount of water present in undesirable places i.e, fields, roads, establishments that are fully or partially submerged in water due to floods, tsunami etc. Thus, these images are kept in a unified Water Disaster category. Similarly, the images with landslides are kept under Land Slide sub-category. The land images with drought are kept in the Drought subcategory. The Urban Fire images tend to have buildings, cars, traffic, and other types of infrastructures with fire whereas the wildfire images normally have trees and other types of greeneries, grasslands and often animals. Damaged Infrastructure category has images where there are broken remnants of buildings or concrete infrastructure, vehicles etc. The structural damages caused by earthquakes are kept under Earthquake subcategory. Human damage category consists of bloody, wounded, burned, and gory pictures due to war or accidents. Bandages and stretchers are also present in some of the images in the human damage category.

During these class label annotation tasks, we have kept all three annotators' labels into account. If all three annotations differ, the image is discarded. If two annotations coincide, the image is put into that category. After this filtration process we have ended up with 13720 images in total.

### C. Creating the training and test sets

For our experiment, we have merged the sub classes of each parent class. Therefore, our training classes are Fire Disaster, Human Damage, Water Disaster, Land Disaster, Damaged Infrastructure, and Non-Damage. We have carefully

handpicked two hundred challenging images and used that as test set. We have tried our best to ensure that these test images reflect the diversity of our dataset.

Our test set consists of images from each parent class. Our primary focus is to keep as much variety as possible in the selected test images. Hence, we have included aerial images, landscapes, low light images, scenarios with many subjects in our test set. We have also put images that would be challenging for the network to classify, like non disaster images with red hue which resemble fire disaster image, sea-beach images which resemble water disaster etc. For each parent class that has sub-classes, we pulled images from each of the sub-classes. We didn't use any random method to collect the test images, rather the selection procedure was carefully performed by human selectors following the above guidelines.

## III. Dataset Characteristics Analysis

### A. Diversity Analysis

One of the key characteristics of our proposed dataset is diversity. But, it is difficult to devise a measure that quantifies diversity. However, to tackle this problem we follow the procedures mentioned in [21]. We compute the average image of each class and measure lossless JPG file size which reflects the amount of information in an image. A diverse image class will result in a blurrier average image, consequently the JPG file size wil be smaller. On the other hand, a less diverse image class will result in a more structured, sharper average image with a greater JPG file size.

### B. Performance and Attention Analysis

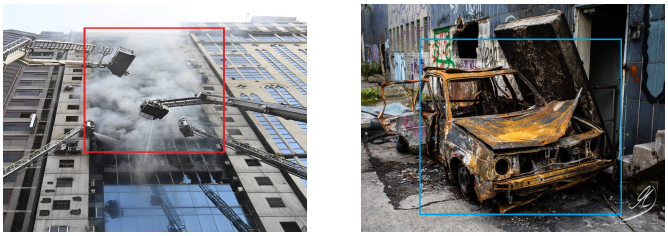
We design our experiments to show the efficacy of our dataset in training deep learning models. Moreover, the experiments are performed to show how the classification models put their attention on particular parts of the images for predicting class label. Our experimental objectives are:

- To measure the quality of the training set for building classifiers. For this purpose, we use five fold cross validation to show the generalization ability across different folds of the dataset.
- To measure the performance of classifiers on unseen data.
- Our final objective is to quantify the attention localization capability of classifiers trained on our dataset.

*1) Classifier Model Description:* We have used VGG-16 as our classification network. The weights of the network are initialized with weights pre-trained on ImageNet data. The input image size is $224 \times 224$. To depict the region of input image where the network is focusing it's attention, two separate attention modules [15], [16] are used:

**Class Activation Map (CAM):** Upon the VGG-16 network, Global Average Pooling (GAP) is used. The outputs of the GAP layer goes to six class softmax layer. To get the attention heatmap the weights of the dominant class is later dot multiplied with the last convolutional layer of VGG-16, which is then upsampled to the input image size.

**Three Layer Attention Map (TLAM):** The TLAM[16] demonstrates how soft trainable attention can improve image classification performance and highlight key parts of images. We use a VGG-16 network where local feature maps $L_i(i = 1, 2, 3)$ are taken from the last three maxpool layers. Then global feature map G is taken after the last conv layer. A parameterized compatibility score is calculated from which attention weights are found. Finally, all three attention maps are concatenated and passed through a fully connected layer to get the final prediction.



(a) Fire Disaster (b) Damaged Infrastructure

Fig. 4: Bounding Box Annotations

*2) Five-fold Cross Validation:* To evaluate the uniformity of the distribution of our dataset, we have performed five fold cross validation. In each validation process, we have selected $80\%$ images from each class for training and rest $20\%$ images for testing. We make sure that no images from this $20\%$ set are selected for testing in other validation process. That means: the training and testing sets are always non-overlapping in each validation iteration.

*3) Testing:* We use both CAM and TLAM architectures for training and then test on the 200 unseen test data that we created. We have reported the classification result for both our dataset and UCI dataset.

We have used a batchsize of 64 and 32 respectively for CAM and TLAM. We use a small learning rate (0.0001) to make sure effective fine tuning occurs. We have used Adam optimizer and weighted cross entropy loss during the training.

*4) Human assessment of visual attention:* To quantify the correctness of focusing attention by deep learning networks trained on our dataset, the test images are annotated by six human annotators. We do this to compare the human way of paying attention with the neural network. The annotation task has been performed following the standard annotations guidelines from the PASCAL Visual Object Classes(VOC) Challenge [22]. Each of our annotators has been trained before the task. Then, they have been asked to draw bounding boxes to the parts of the test images where their attention is intuitively drawn to infer the disaster class. A bounding box is drawn around the disaster region making sure that all of the visible context of the disaster class are tightly inside the bounding box. These bounding box images are later compared with the attention-maps provided by CAM and TLAM to calculate mean Intersection over Union (mIoU). Figure 4 shows the examples of bounding boxes for images from Fire Disaster and Infrastructure Damage classes.

*5) Visualizing Attention and verification:* This experiment is performed to quantify the attention localization capability of classifiers trained on our dataset.

The CAM and TLAM output images are transformed to binarized masks by making normalized attention values greater than a threshold to have intensity value of one and rest of the pixels to zero. The thresholds are 0.15 and 0.10 for CAM and TLAM, respectively. We opted for a lower threshold in case of TLAM because attention heatmap outputs of TLAM are very fine and thin compared to CAM. The annotated test images are also binarized by making pixels in the bounding box to have intensity value of one and the rest to zero. After that, the amount of overlapping between the two masks are calculated using Intersection Over Union (IOU) method. This procedure is performed over all the test images. Then the mean IOU is calculated, which is the quantified score.

*6) Performance Measurement:* We present our classification result in terms of accuracy and macro F1-score. Moreover, we have calculated the mIoU to show how well the classifier's attention overlaps with the human attention.

## IV. Experimental Results and Discussions

In this section we present the results of the experiments that we have designed in last section. We compare our results with UCI dataset to show the efficacy of our proposed dataset.

### A. Diversity

Figure 5 shows the lossless JPG file size of average image for each class of our dataset vs UCI dataset. It is observed that the average images for four out of six classes of our dataset have less byte size and thus contain more information. Therefore, our dataset is more diverse than UCI dataset.

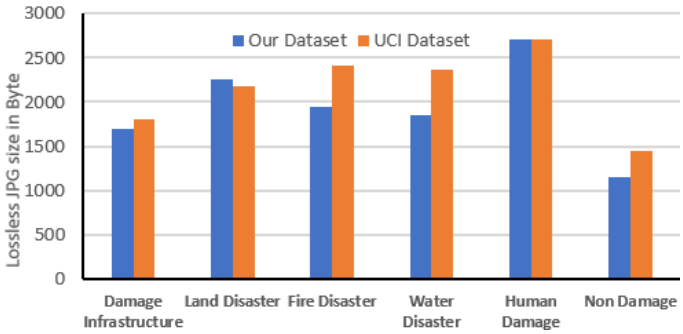



Fig. 5: Lossless JPG size in byte of our dataset vs UCI dataset

### B. Performance and Attention Analysis

*1) Five-fold Cross Validation:* Table III shows the performance of the CAM and TLAM for the 5-fold cross validation experiment. We report accuracy and macro average F1 score for each of the fold tested.

It can be easily observed that the accuracy for each of the fold is close to 0.96 for CAM. Also, the F1-score is around 0.90. Similar results are also observed for TLAM. Accuracy scores are 0.96 and macro F1 scores are around 0.88 for all folds except fold 4. Fold 4 has slightly better performance (accuracy: 0.97 and macro F1 score 0.92). 5-fold cross validation results with both CAM and TLAM suggest that our proposed dataset is well structured and uniform throughout its extent.

TABLE III: Cross Validation Summary for CAM and TLAM

| | CAM | | TLAM | |
|---|---|---|---|---|
| | Accuracy | F1 score (Macro Avg) | Accuracy | F1 score (Macro Avg) |
| Fold 1 | 0.96 | 0.89 | 0.96 | 0.89 |
| Fold 2 | 0.96 | 0.90 | 0.96 | 0.88 |
| Fold 3 | 0.95 | 0.89 | 0.96 | 0.88 |
| Fold 4 | 0.96 | 0.92 | 0.97 | 0.92 |
| Fold 5 | 0.96 | 0.90 | 0.96 | 0.88 |

*2) Testing performance:* In table IV, we report the detailed performance result of test procedure. Discriminating features such as fire, smoke, flame help in better classification of fire images. Moreover, as the Water Disaster images of our dataset have unique characteristics, the F1 score is high for this class.

Table IV clearly shows the efficacy of our dataset in training the classifier models. The test images are classified with significantly high F1-score. The macro average F1 score for CAM and TLAM are: 0.96 and 0.97, respectively. In contrast the macro F1 scores of the same classifiers are very low: 0.42 and 0.39 when the classifiers are trained with UCI dataset. Most importantly, the UCI dataset cannot make the classifiers learn effective discriminative features for Human Damage, Fire Disaster, and Land Disaster. We observed that for most of the test images, the classifiers have put attentions completely in the wrong regions.

Figure 6 presents some samples of attention heatmaps. In this figure, first row has the input images, middle row contains examples of misplaced attention. Third row shows that attention is moved to the correct region when we train CAM with our dataset. The heatmap represents the attention intensity. Reddish heatmap indicates higher attention.

6(a) is a test image from Fire Disaster class. But UCI trained CAM model has put it's attention on the road as seen in the image 6(g). In contrast, CAM trained with our dataset learns to pay it's attention correctly on the smoke region as shown in 6(m).

A test image from Water Disaster class is in 6(b) and the attention heatmap is shown in 6(h) when CAM is trained with UCI. It is easy to see that classifier has put it's attention on the vehicles and human resulting in wrong classification. The image 6(n) shows that the attention is in the expected water region; thus the CAM model could correctly classify it as Water Disaster image.

Similarly, 6(d) is a Draught image that falls under the Land Disaster class. UCI trained CAM pays attention towards the human in 6(j) whereas the CAM model trained with our dataset puts attention on the dry and fractured land in 6(j) to classify the image as Land Disaster.

We have also devised an experiment to show how well our dataset generalizes on UCI dataset. As UCI dataset does not have any explicit test set, we randomly pick 40 images from each class and make a UCI test set of 240 images.

TABLE IV: Performance Summary for CAM and TLAM on Test Data

| | CAM trained on proposed training set | | CAM trained on UCI dataset | | TLAM trained on proposed training set | | TLAM trained on UCI dataset | |
|---|---|---|---|---|---|---|---|---|
| | Precision | F1 | Precision | F1 | Precision | F1 | Precision | F1 |
| Infrastructure Damage | 0.91 | 0.93 | 0.71 | 0.81 | 0.87 | 0.92 | 0.67 | 0.78 |
| Fire Disaster | 1.00 | 0.98 | 0.03 | 0.03 | 1.00 | 1.00 | 0.00 | 0.00 |
| Human Damage | 0.91 | 0.92 | 0.00 | 0.00 | 0.94 | 0.97 | 0.00 | 0.00 |
| Water Disaster | 1 | 1.00 | 0.91 | 0.94 | 1.00 | 0.98 | 0.94 | 0.94 |
| Land Disaster | 0.94 | 0.94 | 0.06 | 0.06 | 1.00 | 0.94 | 0.00 | 0.00 |
| Non Damage | 0.97 | 0.96 | 1.00 | 0.65 | 1.00 | 0.99 | 0.70 | 0.61 |
| Marco Average | 0.96 | 0.96 | 0.45 | 0.42 | 0.97 | 0.97 | 0.38 | 0.39 |

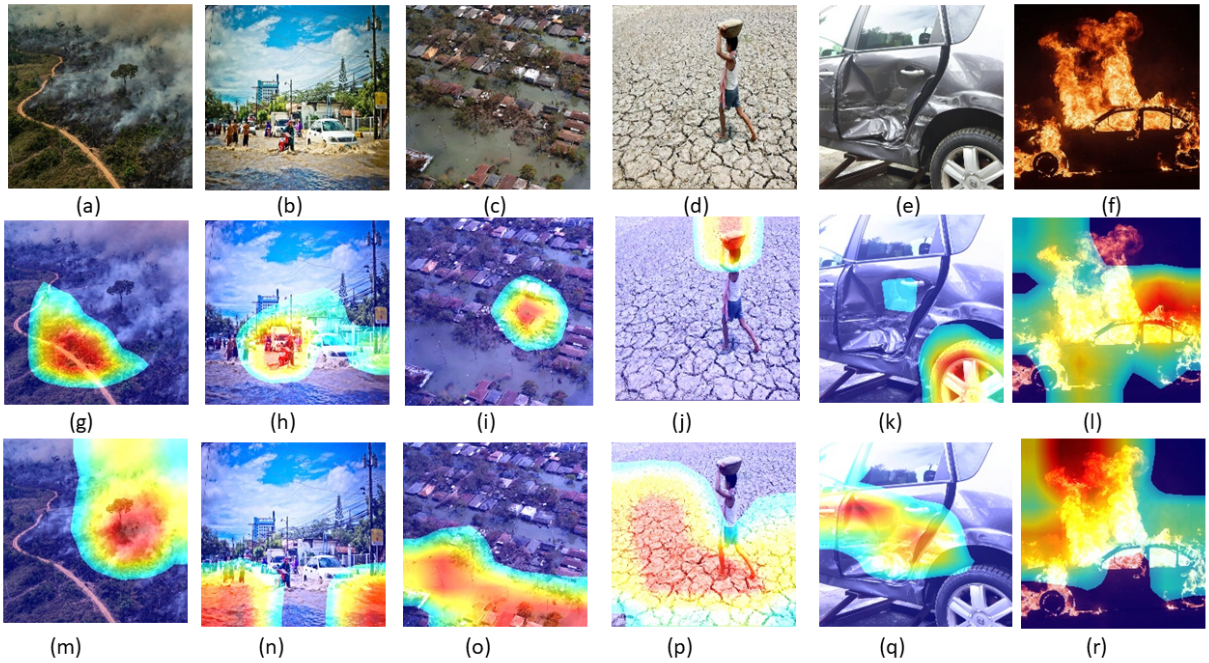



Fig. 6: Top row: Input images from (a) Fire Disaster, (b) Water Disaster, (c) Water Disaster, (d) Land Disaster, (e) Infrastructure Damage, (f) Fire Disaster; Middle row: Wrong attention (Using CAM trained with UCI dataset); Last row: Correct attention (Using CAM trained with our dataset)

Then we compare the classification performance on UCI test set using CAM model trained on both our proposed dataset and rest of the UCI dataset. To make the result unbiased, we perform the random picking and testing five times. The average classification accuracy of the five runs of testing on UCI test set is 71.25% with CAM trained on our dataset and 68.62% with CAM trained on UCI dataset. So, our dataset generalizes well on UCI dataset.

*3) BOWFIRE:* To further show the generalization capability, we have evaluated our model on benchmark disaster dataset. We have used the Bowfire test set [7] that has only two classes: Fire and No-Fire. We have tested the Bowfire test set with CAM and TLAM that are trained with the proposed dataset. We observe that the F1 score of fire class with CAM (trained with our dataset) is $0.84$, whereas, the F1 score reported in the paper [23] for [7] is in the range of $0.6-0.7$ and for [24] it is in range $0.5-0.6$. When we have performed the experiment with TLAM (trained with our dataset), we have got exactly the same F1 score $0.84$.

*4) Classifier's Attention vs Human Attention:* In order to show the quality of our dataset, we calculate the mean Intersection over Union (mIoU) of human attention and classifier's attention for our test data while the classifier has been trained with our training set. Additionally, we calculate mIou on test set for classifier trained with UCI dataset. Table V shows that the mIoU is significantly higher when CAM has been trained with our training set. The mIoU for CAM trained on our proposed dataset is $0.53$ whereas the mIoU drops to $0.45$ for CAM trained with UCI dataset. After doing the same experiment with TLAM, the mIoU significantly drops from $0.31$ to $0.18$.

To show how subcategorization empowers classifiers' attention, we have trained CAM with all images except the images from Wildfire subcategory. The test mIoU drops from 0.53 to

0.5. Furthermore, we again train a new CAM model discarding Wildfire and Drought images. The test mIoU then drops to 0.49.

Apart from CAM and TLAM, we have also experimented with recent attention module GradCam++ [25]. We train using resnet-101 architecture on both our dataset and UCI dataset. We then compare the overlap agreement between human attention and classifier's attention. Our dataset yields an mIOu of 0.56 and for UCI it is 0.49. So, GradCam++ also shows the efficacy of our dataset.

The experimental results suggest that the structure of the proposed dataset is such well organized and diverse that attention localization capability of classifiers are much improved.

TABLE V: Overlap agreement between human attention and classifier attention

| | Human-Classifier Attention Overlap (Mean IoU) | Human-Classifier Attention Overlap (Mean IoU) |
|---|---|---|
| Training Set | Proposed Dataset | UCI Dataset |
| CAM | 0.53 | 0.45 |
| TLAM | 0.31 | 0.18 |

## V. Conclusion

Accumulating a comprehensive image dataset for disaster detection task is very challenging, especially, images that contain representative information for different classes. Also, it is a difficult task to provide annotation for such images. To reduce biases in the classification process we have provided class information for each image by three different individuals. Also six annotators provided bounding box annotations for test images. In future, we plan to provide bounding box ground truth annotations for all the images of our dataset. More challenging issue is to come up with a proper assessment mechanism for such datasets. It is often not enough to look only at the quantitative measures such as classification accuracy, precision, recall, F1 score, etc. In this work we try analysing the performance of the attention based classifiers not only to show that, systems trained with our dataset can outperform exact same systems trained with other dataset, moreover, we show that visually and numerically, human level attention can be achieved if attention based classifiers are trained with our dataset.

## VI. Acknowledgment

This project is supported by ICT Division, Government of Bangladesh, and Independent University, Bangladesh (IUB).

## References

[1] Arif, A. Omar, S. Ashraf, A. M. Rahman, M. A. Amin, and A. A. Ali, "A comparative study on disaster detection from social media images using deep learning," in *Global AI Congress*, 2019.

[2] Y. Rizk, H. S. Jomaa, M. Awad, and C. Castillo, "A computationally efficient multi-modal classification approach of disaster-related twitter images," in *Proceedings of the 34th ACM/SIGAPP symposium on applied computing*, 2019, pp. 2050–2059.

[3] P. Giannakeris, K. Avgerinakis, A. Karakostas, S. Vrochidis, and I. Kompatsiaris, "People and vehicles in danger - a fire and flood detection system in social media," 06 2018.

[4] K. Muhammad, J. Ahmad, and S. Baik, "Early fire detection using convolutional neural networks during surveillance for effective disaster management," *Neurocomputing*, 12 2017.

[5] B. C. Ko, S. J. Ham, and J. Y. Nam, "Modeling and formalization of fuzzy finite automata for detection of irregular fire flames," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 12, pp. 1903–1912, 2011.

[6] S. Verstockt, T. Beji, P. De Potter, S. Hoecke, B. Sette, B. Merci, and R. Van de Walle, "Video driven fire spread forecasting (f) using multi-modal lwir and visual flame and smoke data," *Pattern Recognition Letters*, vol. 34, pp. 62 – 69, 01 2013.

[7] D. Chino, L. Avalhais, J. Rodrigues Jr, and A. Traina, "Bowfire: Detection of fire in still images by integrating pixel color and texture analysis," 08 2015, pp. 95–102.

[8] P. Foggia, A. Saggese, and M. Vento, "Real-time fire detection for video-surveillance applications using a combination of experts based on color, shape, and motion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 9, pp. 1545–1556, 2015.

[9] F. Alam, M. Imran, and F. Ofli, "Image4act: Online social media image processing for disaster response," 07 2017, pp. 601–604.

[10] H. Mozannar, Y. Rizk, and M. Awad, "Damage identification in social media posts using multimodal deep learning," 05 2018.

[11] Disaster dataset. [Online]. Available: https://drive.google.com/open?id=1VvkBRIYW6oD31K3gkPk4-4nlGE2poXFU

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[13] K. Avgerinakis, A. Moumtzidou, S. Andreadis, E. Michail, I. Gialampoukidis, S. Vrochidis, and Y. Kompatsiaris, "Visual and textual analysis of social media and satellite images for flood detection @ multimedia satellite task mediaeval 2017," in *MediaEval*, 2017.

[14] D. T. Nguyen, F. Ofli, M. Imran, and P. Mitra, "Damage assessment from social media imagery data during disasters," in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, 2017, pp. 569–576.

[15] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*.

[16] S. Jetley, N. A. Lord, N. Lee, and P. H. Torr, "Learn to pay attention," *arXiv preprint arXiv:1804.02391*, 2018.

[17] J. Holmes and K. Sherin. (October 28, 2019) 40 photos show the incredible destruction wrought by the 2019 california wildfires. [Online; accessed 23-January-2020]. [Online]. Available: www.esquire.com/news-politics/g29610550/california-wildfire-photos-2019/?slide=9

[18] Pictures from the amazon rainforest fires. [Online; accessed 25-January-2020]. [Online]. Available: https://www.cbsnews.com/pictures/pictures-amazon-rainforest-fires-in-brazil/17

[19] . (August 18, 2018) Yemen conflict: Un experts detail possible war crimes by all parties. [Online; accessed 28-January-2020]. [Online]. Available: https://www.bbc.com/news/world-middle-east-45329220

[20] J. B. Puneet Bansal, "Intel image classification," January 2019. [Online]. Available: https://www.kaggle.com/puneet6060/intel-image-classification

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[22] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[23] K. Muhammad, J. Ahmad, and S. W. Baik, "Early fire detection using convolutional neural networks during surveillance for effective disaster management," *Neurocomputing*, vol. 288, pp. 30–42, 2018.

[24] T. Celik and H. Demirel, "Fire detection in video sequences using a generic color model," *Fire Safety Journal*, vol. 44, 2009.

[25] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 839–847.