Damage Assessment from Social Media Imagery Data During Disasters

Conference Paper · July 2017

DOI: 10.1145/3110025.3110109

4 authors:

Dat T. Nguyen

6 PUBLICATIONS **84** CITATIONS

SEE PROFILE

Ferda Ofli

Qatar Computing Research Institute

75 PUBLICATIONS **2,071** CITATIONS

SEE PROFILE

Muhammad Imran

Qatar Computing Research Institute

89 PUBLICATIONS **2,281** CITATIONS

SEE PROFILE

Prasenjit Mitra

Pennsylvania State University

299 PUBLICATIONS **6,435** CITATIONS

SEE PROFILE

Damage Assessment from Social Media Imagery Data During Disasters

Dat T. Nguyen*, Ferda Ofli*, Muhammad Imran*, Prasenjit Mitra**
*Qatar Computing Research Institute, HBKU, Doha, Qatar.
**The Pennsylvania State University, University Park, PA, USA
{ndat,fofli,mimran}@hbku.edu.qa; pmitra@psu.edu

***Abstract*—Rapid access to situation-sensitive data through social media networks creates new opportunities to address a number of real-world problems. Damage assessment during disasters is a core situational awareness task for many humanitarian organizations that traditionally takes weeks and months. In this work, we analyze images posted on social media platforms during natural disasters to determine the level of damage caused by the disasters. We employ state-of-the-art machine learning techniques to perform an extensive experimentation of damage assessment using images from four major natural disasters. We show that the domain-specific fine-tuning of deep Convolutional Neural Networks (CNN) outperforms other state-of-the-art techniques such as Bag-of-Visual-Words (BoVW). High classification accuracy under both event-specific and cross-event test settings demonstrate that the proposed approach can effectively adapt deep-CNN features to identify the severity of destruction from social media images taken after a disaster strikes.**

I. Introduction

The use of microblogging social networks such as Twitter has become widespread, especially, during mass emergencies such as natural or man-made disasters. People use social networks to post situational updates in a variety of forms such as textual messages, images, and videos [1], [2]. Many studies have shown that this online information is useful for rapid crisis response and management [1], [3]. For example, after major disasters, researchers use Twitter data to find out about the number of injured or dead people, most urgent needs of affected people (e.g. shelter, food, water, etc.), donation offers, etc. [4], [5].

Artificial Intelligence for Disaster Response (AIDR) is a system conceived and developed at Qatar Computing Research Institute (QCRI) to harness information from real-time tweets that emerge from an area struck by a natural disaster to help coordinate relief activities [6]. The AIDR system combines human and machine intelligence to categorize crisis-related messages during the sudden-onset of natural or man-made disasters. AIDR is now routinely used by the United Nations Office for the Coordination of Humanitarian Affairs (UN OCHA) and many other emergency departments in the world. The system has been used during several major disasters such as the 2015 Nepal Earthquake, Typhoon Hagupit by a number of different humanitarian organizations including UN OCHA and UNICEF. The AIDR system was originally designed to process textual content in real-time.

At the onset of a crisis, assessment of the level of damage is one of the key situational awareness requirements of humanitarian response organizations to understand the severity of destruction and to plan relief efforts accordingly. It is important for first-responders to know what type of damage happened and where. Existing research on the use of Twitter during an emergency for damage assessment is mainly focused on the textual content of tweets [7]. Despite the recent advances in computer vision field, especially in image classification, most of the existing works for emergency management do not yet rely on the image data.

However, in this work, we propose an extension of the AIDR system to process imagery data posted on social networks during disasters for damage assessment [8]. Specifically, our goal is to determine whether the assessment of the severity of damage through images is possible or not. For damage assessment, we consider three levels: *severe damage, mild damage,* and *little-to-no damage* (further details regarding each damage level are given in Section III). Given the fact that, during disasters, tens of thousands of images are posted on social media platforms such as Twitter, a simple automatic damage assessment system to learn whether an image contains damage or not will not be considered that helpful for emergency responders than a system that not only detects damage-related images but also determines the level of damage (i.e., severe, mild, or low). This will greatly help emergency responders prioritize their relief efforts and planning for the most severe cases first. To the best of our knowledge, no prior work has dealt with the task of identifying level of damage (i.e., severe, mild, or low) from social media images as our work does.

Analyzing large volumes of images generated after a major disaster remains a challenging task in contrast to the ease of acquiring them from social media platforms due to the following reasons:

ASONAM '17, July 31 - August 03, 2017, Sydney, Australia

ACM ISBN 978-1-4503-4993-2/17/07?/$15.00
http://dx.doi.org/10.1145/3110025.3110109

A. Inherent task complexity

State-of-the-art methods in computer vision leverage on large image collections with clean annotations of well-defined object categories such as house, car, airplane, cat, dog, etc. However, images acquired after major disasters do not have the same characteristics as those used in the traditional computer vision research. In disaster images, objects are generally not in well-defined forms, e.g., uprooted trees, damaged buildings, destroyed roads and bridges (for example see the severe and mild damage images of the Nepal and Ecuador Earthquakes shown in Figure 1) Hence, disaster image analysis requires more robust algorithms to operate on *image datasets in the "extreme" wild.*

B. Poor signal-to-noise ratio:

Even though the volume of images collected from social media platforms on the wake of disasters can be large, the level of noise in the resulting datasets is also extremely high. Therefore, the amount of labeled imagery *relevant to the actual disaster event* is relatively small with respect to the large image collections used in the traditional computer vision research with less noise [8].

C. Task subjectivity:

The terms "severe", "mild" or "no" damage are subjective terms. Humans find it difficult to always agree on whether the damage is severe or mild, or even in some cases whether the damage is mild or none. Besides, the scope of the damage is taken into account differently by different annotators. For example, if in an image, one building out of, say 100, is destroyed, does that image qualify as depicting "severe" damage? Or because the damage is restricted to a small part of the image, should it be tagged as "mild"? Given this inherit complexity of the damage assessment task, the quality of the annotations obtained from the crowd is therefore lower than the quality of annotations available in the traditional large-scale image datasets due to high *disagreement between the annotators on the severity of damage*. For instance, for many human-annotators the image of a partially destroyed home of the Nepal Earthquake shown in Figure 1 (Nepal Mild) was a complex task, as according to some annotators it was a mild damage, whereas, some other annotators disagreed and annotated it as severe damage.

D. Cold-start issue:

The first few hours of a disaster are critical for response organizations and this is the time-period when information is usually missing. Relying on traditional information collection approaches like sending human analysts to the disaster zone or relying on main-stream media could take several days or weeks. Social media can be very helpful during this information blackout period. However, enabling machines to start the damage assessment task from images requires labeled data (i.e. human-annotated images). To acquire an adequate number of labeled data at the onset of a disaster event is not a trivial task either due to limited resources or less budget. Often this process takes several hours thus delays machine training and prediction tasks. To overcome this issue, in this work, we aim to test whether labeled data from past events as well as from the Web can be used to predict images of a target event [9].

Given these challenges, our objective in this study is to explore the performance of several image classification techniques for the desired task of tagging each given disaster image collected from social media platforms with a label indicating the level of damage observed in that particular image. For this purpose, we employ both traditional computer vision techniques such as Bag-of-Visual-Words (BoVW) model as well as state-of-the-art deep learning techniques such as Convolutional Neural Networks (CNN) to assess the level of damage in disaster images.

Our experimental analyses show that domain-specific fine-tuning of deep CNNs outperforms the traditional BoVW models by a considerable margin. Moreover, leveraging on labeled data from past disaster events as well as data collected from the Web help to address the *cold-start* problem at the wake of a new disaster.

The rest of the paper is organized as follows. We provide a brief literature review in Section II. We describe our disaster image datasets and human annotations in Section III. In Section IV, we elaborate on the learning schemes and settings used in our experiments. We present our experimental results in Section V. Discussion about the results, limitations, and potential future directions are discussed in section VI. Finally, we conclude the paper in Section VII.

II. Related work

State-of-the-art methods in computer vision domain leverage on large labeled image collections such as PASCAL VOC [10] or ImageNet [11] by using various deep learning architectures based on convolutional neural networks (CNN) [12]–[16]. For example, the winner of the 2016 ILSVRC challenge has achieved as low as 2.99% top-5 classification error with an ensemble method based on existing deep CNN architectures for image classification such as Inception networks [17], Residual Networks [16] and Wide Residual Networks [18]. These deep CNNs integrate low, medium and high level features [19] and classifiers in an end-to-end fashion to optimize on the target prediction task directly from raw data. For example, the lower layers of deep CNN architectures correspond to a representation suitable for low-level vision tasks while the higher layers are more domain specific [20] and eliminate the need for pre-defined feature engineering like Scale Invariant Feature Transform (SIFT) [21] and Histogram of Oriented Gradients (HOG) [22].

Prior to deep-CNN era, the most popular techniques were based on Bag-of-Visual-Words (BoVW) models [23], [24] where handcrafted local image features such as SIFT were first extracted and vector quantized into a visual vocabulary by k-means clustering. And each image was then represented by a histogram of how often the extracted local features were assigned to each visual word (hence the name "bag-of-visual-words"). A support vector machine classifier was eventually

used on top of these histogram representations to perform the classification task. Later on, classification-by-detection type of approaches became popular. Among others, Deformable Part-based Models (DPM) [25] and its variations [26], [27] were the state-of-the-art techniques before the revitalization of deep-CNN architectures for image classification and object detection.

On the other hand, use of computer vision and image processing techniques for damage assessment from images has not been fully explored in literature, except for a handful of studies that has sporadically appeared mainly in the remote sensing domain which were mostly based on analyses of satellite [28], [29] and aerial [30]–[32] imagery. Most recent studies in this direction, and also the most relevant ones to the current study, are presented by Lagerstrom et al. [33] and by Daly and Thom [34] where both studies analyzed social media data but in a binary image classification setting for fire/not-fire detection scenario. In our current work, we address a more challenging problem since we neither limit the task to a particular disaster type nor to a binary classification setting.

III. Datasets and Human Annotations

A. Dataset Collection

In this study, we leveraged on labeled data from past disaster events as well as data collected from the Web.

1) Disaster-specific image data: We used the AIDR platform [6] to collect images from social media platforms such as Twitter during four major natural disasters, namely, Typhoon Ruby/Hagupit, Nepal Earthquake, Ecuador Earthquake, and Hurricane Matthew. The data collection was based on event-specific hashtags and keywords. Table I lists all the datasets and total number of images initially collected in each dataset. Figure 1 shows example images from our datasets.

2) Google image data: Often the scarcity of labeled data (e.g. labeled images in our case) at the onset of a crisis situation causes a delay in building machine learning classifiers if we depend upon labeled images from the event to train the model. To overcome this cold-start issue, in this work, we aim to use damage-related images from the Web to determine if the image data on the Web can help train an effective classifier for damage assessment. We use Google search to query damage-related images. Specifically, we used queries like *damage building, damage bridge, damage road* to crawl such images.

B. Human Annotations

We acquire human labels to train and evaluate our models. For this purpose, we get labels from two different settings. The first set of labels were gathered from AIDR. In this case, volunteers, during a crisis situation, are employed to help label images. In the second setting, we use the Crowdflower[1] crowdsourcing platform to annotate images. The following instructions were given to the annotators for the image labeling tasks.

[1]http://crowdflower.com/

TABLE I
Dataset details for all four disaster events and images collected from Google.

Disaster Name	Year	Number of Images
Typhoon Ruby/Hagupit	2014	7,000
Nepal Earthquake	2015	57,000
Ecuador Earthquake	2016	65,000
Hurricane Matthew	2016	61,000
Google Images	NA	20,000

Damage severity levels instructions:
The purpose of this task is to assess the severity of damage shown in an image. The severity of damage in an image is the extent of physical destruction shown in it. We are only interested in structural damages like broken bridges, collapsed or shattered buildings, destroyed or cracked roads etc. An example of a non-structural damage is where we can see smoke due to fire on a building or a bridge. In this particular task, we do not consider such non-structural damage types. So in such cases, the annotators were asked to select the "no damage" category.

1- Severe damage: Images that show substantial destruction of an infrastructure belong to the severe damage category. A non-livable or non-usable building, a non-crossable bridge, or a non-drivable road are all examples of severely damaged infrastructures.

2- Mild damage: Damage generally exceeding minor [damage] with up to 50% of a building, for example, in the focus of the image sustaining partial loss of amenity/roof. Maybe only part of the building has to be closed down, but other parts can still be used. In case of a bridge, if the bridge can still be used, but, part of it is unusable and/or needs some amount of repairs.

3- Little-to-no damage: Images that show damage-free infrastructure (except for wear and tear due to age or disrepair) belong to the no-damage category.

Given the aforementioned instructions for crowdsourcing, we used the AIDR platform during the actual disasters to obtain volunteers-based annotations. Moreover, we obtained more annotations using the Crowdflower platform after the disasters events were over. To maintain high-quality, each crowd task required an agreement of at least three different workers to finalize a label for an image. Table II shows crowdsourcing results in terms of labeled images for each dataset annotated using AIDR and Crowdflower.

IV. Experimental Setting

In this section, we first elaborate on several image classification techniques employed in our experiments, and then describe different experimental settings explored in our study.

A. Learning Schemes

1) Bag-of-Visual-Words model: State-of-the-art techniques in pre-deep learning era relied on Bag-of-Visual-Words

Fig. 1. Sample images with different damage levels from different disaster datasets.

(BoVW) models that used handcrafted features such as Scale Invariant Feature Transform (SIFT) or Histogram of Oriented Gradients (HOG) together with classical machine learning classifiers such as Support Vector Machines (SVM) or Random Forests (RF). In this study, we extracted Pyramidal Histogram of visual Words (PHOW) features (a variant of dense multi-scale SIFT descriptors proposed in [35]) from each image[2]. We used pyramid histogram of N visual words with 3 levels to extract visual features from images. At level 0, each image was divided into a 2×2 grid, yielding a total of 4 regions. At level 1, each region was represented by a $2N$-vector and at level 2 by a $1N$-vector. Consequently, we obtained BoVW vectors with $12 \times N$ dimensions. Since our spatial histograms had $N = 300$ visual words, we ended up with 3600-dimensional vector representations for each image. We then trained a linear SVM classifier on top of these features to perform the multi-class classification task. We consider the resulting BoVW model as our baseline technique as it was also recently used in the related work by Lagerstrom et al. [33].

[2]We used VLFeat toolkit available at http://www.vlfeat.org.

2) Pre-trained CNN as feature extractor: CNNs are rarely trained from scratch for new datasets because state-of-the-art CNNs (*i*) are getting deeper everyday, and (*ii*) require larger datasets to train on. However, collecting large datasets for the particular problem at hand is usually hard in practice (as in the current study). Therefore, it is common to devise new techniques based on pre-trained networks. A popular approach is to use a pre-trained CNN as a feature extractor on the new dataset. That is, an image from the new dataset is simply fed as input to the pre-trained network and different layers of the pre-trained network are used as feature extractors where each image can be mapped to a new representation. In this study, we used the VGG16 network trained on the ImageNet dataset using over 1.2M images and 1000 categories [11]. The VGG16 network consists of 16 layers and around 140 million weight parameters [14]. We opted to use fc7 layer as our 4096-dimensional deep feature extractor (denoted as VGG16-fc7) in our experiments. Features were computed by forward propagating a mean-subtracted 224×224 RGB image through thirteen convolutional and two fully connected layers.

TABLE II
NUMBER OF LABELED IMAGES FOR EACH DATASET. IMAGES TAGGED BY CROWDFLOWER ARE DENOTED AS (CF) WHILE REST OF THE TAGS ARE OBTAINED FROM AIDR [6].

Classes	Nepal Earthquake	Ecuador Earthquake	Typhoon Ruby	Hurricane Matthew	Google
Severe	8,927	844 + 390(CF)	91	112 + 16(CF)	1130 (CF)
Mild	2,257	89 + 45(CF)	342	94 + 48(CF)	887 (CF)
None	7,920	791 + 121(CF)	400	127 + 199(CF)	990 (CF)
Total	19,104	2,218	833	596	3,007

3) Fine-tuning a pre-trained CNN: Another popular approach is to use the existing weights of a pre-trained CNN as an initialization for the new dataset, which is often referred to as fine-tuning. In this transfer-learning approach, the last layer of the network is adapted to the task at hand (i.e., number of categories in the softmax layer and sometimes even the loss function) and the pre-trained network is fine-tuned according to the training images from the new dataset. In this study, we used a cross-entropy loss function defined as follows:

$$J(\theta) = -\sum_{n=1}^{N} y_n \log(\hat{y}_{n\theta}) + \gamma \sum_{\theta} \theta^2 \quad (1)$$

where y_n represents one-hot vector of labels and $\hat{y}_{n\theta}$ is the predicted class probabilities by the model for n-th training example in a batch of N images, γ is the multiplier of the L_2 regularization term and θ are the model parameters.

This approach allows us to transfer the features and the parameters of the network from the broad domain (i.e., large-scale image classification) to the specific one (i.e., disaster image analysis). For this purpose, we adapted the VGG16 network pre-trained on the ImageNet dataset to classify the images in our disaster image dataset into one of the three damage categories, i.e., severe, mild, and little-to-no damage.

B. Learning Settings

For fine-tuning experiments, we used back-propagation with minibatch stochastic gradient descent with momentum. We used a batch size of 128 and a momentum of 0.9. We set the regularization multiplier as 5×10^{-4} and drop-out ratio as 0.5. We started fine-tuning with a learning rate of 10^{-3} and decreased it by a factor of 10 after 3 epochs. In total, the learning rate was decreased 2 times, and fine-tuning stopped after 9 epochs since we observed that increasing the number of epochs resulted in overfitting. Before running the experiments, we divided each dataset into three subsets: training (60%), development (20%), and test (20%).

While training a classifier using any of the techniques described above, we consider different real-world scenarios. These scenarios are mainly motivated based on the availability of training data either from the event for which the prediction task needs to be carried out or the training data taken from past events. when no event-specific labeled data is available. This work mainly examines the performance of a classifier when trained and tested in two different settings, as explained below.

1) Event-specific setting: In the event-specific setting, the training, development, and test sets comprised of the labeled data from the same event for which the prediction task was carried out. For instance, to predict the level of damage from Nepal Earthquake images, in this setting, we train a classifier using the 60% of Nepal labeled data and test the trained model on the 20% of the Nepal data. If the event-specific labeled data is available, this particular training setting is ideal to achieve good classification accuracy.

2) Cross-event setting: In the cross-event setting, the basic assumption is that the event, for which the prediction is required, has very little or no labeled data. In this case, we use labeled data from past events to determine its usefulness. This is an important scenario, as in many cases, either due to limited resources or less budget, obtaining event-specific labeled data is not possible or expensive. We perform several cross-event experiments in which we train models on earthquakes and typhoons separately. Specifically, we consider the Ecuador Earthquake and Typhoon Ruby as test events, i.e. we perform predictions on these two events. For instance, to run a prediction task on the Ecuador event data, we train four different models (*i*) based only on the 60% of the Google labeled data (*ii*) based only on the 60% of the Nepal Earthquake data, (*iii*) by combining the 60% of Nepal and the 60% of Ecuador, and (*iv*) by combining the 60% of Nepal and 60% of Ecuador with 60% of Google labeled data. Similarly, we performed four cross-event experiments for Typhoon Ruby as our test event while investigating similar training set combinations of the Typhoon Ruby, Hurricane Matthew and Google labeled data.

The intuition behind using the Google data is when neither event-specific nor past events labeled data is available, we use image data collected from Google to train classifiers. With this setting, we aim to determine the usefulness of Google image data when no other disaster-specific labeled data sources are available.

V. RESULTS

Table III shows the macro-averaged results in terms of precision, recall, F1 score, and accuracy of all the event-specific experiments using the three learning techniques (i.e. BoVW, VGG16-fc7, and VGG16-fine-tuned). As described in

TABLE III
EVENT-SPECIFIC MACRO-AVERAGED RESULTS FOR ALL EVENTS IN TERMS OF PRECISION, RECALL, F1 SCORE, AND OVERALL ACCURACY USING THREE DIFFERENT LEARNING SCHEMES.

	Nepal Earthquake				Ecuador Earthquake			
	Acc.	Precision	Recall	F1	Acc.	Precision	Recall	F1
BoVW	0.78	0.77	0.78	0.77	0.82	0.81	0.82	0.81
VGG16-fc7	0.76	0.76	0.76	0.76	0.82	0.82	0.82	0.82
VGG16-fine-tuned	**0.84**	**0.82**	**0.84**	**0.82**	**0.87**	**0.86**	**0.87**	**0.86**
	Hurricane Matthew				Typhoon Ruby			
	Acc.	Precision	Recall	F1	Acc.	Precision	Recall	F1
BoVW	0.64	0.64	0.66	0.64	0.73	0.74	0.73	0.72
VGG16-fc7	0.63	0.63	0.64	0.63	0.79	0.80	0.80	0.80
VGG16-fine-tuned	**0.74**	**0.73**	**0.74**	**0.74**	**0.81**	**0.80**	**0.81**	**0.80**
	Google Image							
	Acc.	Precision	Recall	F1				
BoVW	0.57	0.53	0.56	0.54				
VGG16-fc7	0.60	0.63	0.64	0.63				
VGG16-fine-tuned	**0.67**	**0.67**	**0.67**	**0.67**				

TABLE IV
CROSS-EVENT MACRO-AVERAGED RESULTS OBTAINED USING THE CNNS WITH FINE-TUNING TECHNIQUE WHEN TRAINED ON 60% OF A GIVEN EVENT DATA (ROWS) AND TESTED ON 20% OF ECUADOR IN ALL CASES.

Training Set Combination	Acc.	Precision	Recall	F1
Google only	0.64	0.77	0.64	0.65
Nepal only	0.82	0.81	0.82	0.88
Nepal + Ecuador	0.88	0.87	0.88	0.87
Google + Nepal + Ecuador	**0.90**	**0.89**	**0.90**	**0.89**

TABLE V
CROSS-EVENT MACRO-AVERAGED RESULTS OBTAINED USING THE CNNS WITH FINE-TUNING TECHNIQUE WHEN TRAINED ON 60% OF A GIVEN EVENT DATA (ROWS) AND TESTED ON 20% OF MATTHEW IN ALL CASES.

Training Set Combination	Acc.	Precision	Recall	F1
Google Only	0.49	0.53	0.49	0.48
Ruby Only	0.68	0.76	0.68	0.69
Ruby + Matthew	0.75	**0.77**	0.75	0.75
Google + Ruby + Matthew	**0.76**	**0.77**	**0.77**	**0.76**

the previous sections, in the event-specific setting, the training and test examples are drawn from the same event data, so ideally, if enough training examples are available to train a classifier then this setting should perform better than all others. Considering BoVW and VGG16-fc7 as our baselines, we can clearly see that our proposed approach (i.e. VGG16 with fine-tuning) outperforms both baseline techniques. Figure 2 shows the precision-recall curves and AUC obtained from the four disaster event models. In the cases of Nepal Earthquake and Hurricane Matthew, we can observe that the *mild damage* class as the most difficult class to learn compared to the *severe* and *none* classes. This finding is inline with what we also observed in the human-labeling results (using inter-annotator agreement), i.e., most of the times humans also face difficulties distinguishing between mild and severe damage images. However, in the case of Typhoon Ruby, we observe low performance on the *severe damage* class. We believe this poor performance is due to the low prevalence of the severe class (i.e., only 91 severe images) in the Ruby labeled data.

In the next experiments, we only use the VGG16 with fine-tuning approach to train classifiers, as it is the best-performing learning scheme in all the event-specific experiments. Table IV shows cross-event macro-averaged results when the Ecuador data was considered as the test event. We trained four different models using the VGG16 with the fine-tuning technique using different combinations of training data. We can see in all the cases, the model trained on the combination of (Google + Nepal + Ecuador) training data outperforms. Interestingly, the Google data helps achieve good performance when combined with other disaster events data compared to when used alone. Figure 3 shows precision-recall curves and AUC for the best model, i.e., Google + Nepal + Ecuador. Clearly, the classifier shows poor performance classifying the mild damage class. We observe almost same results from the model trained on Google + Ruby + Matthew. The overall accuracy of this model, as shown in Table V, is also low. Overall, we consider less training data and class ambiguity between mild and severe classes as the two main issues causing classifiers to poorly perform on the mild damage class.

VI. DISCUSSION

Although social media data is useful in the first few hours after a disaster hit, processing and making-sense out of it is a challenging task. This work aimed at analyzing imagery data collected during a number of disasters for the assessment of the severity of damage to different types of infrastructures. This could be highly useful for response organizations to plan relief

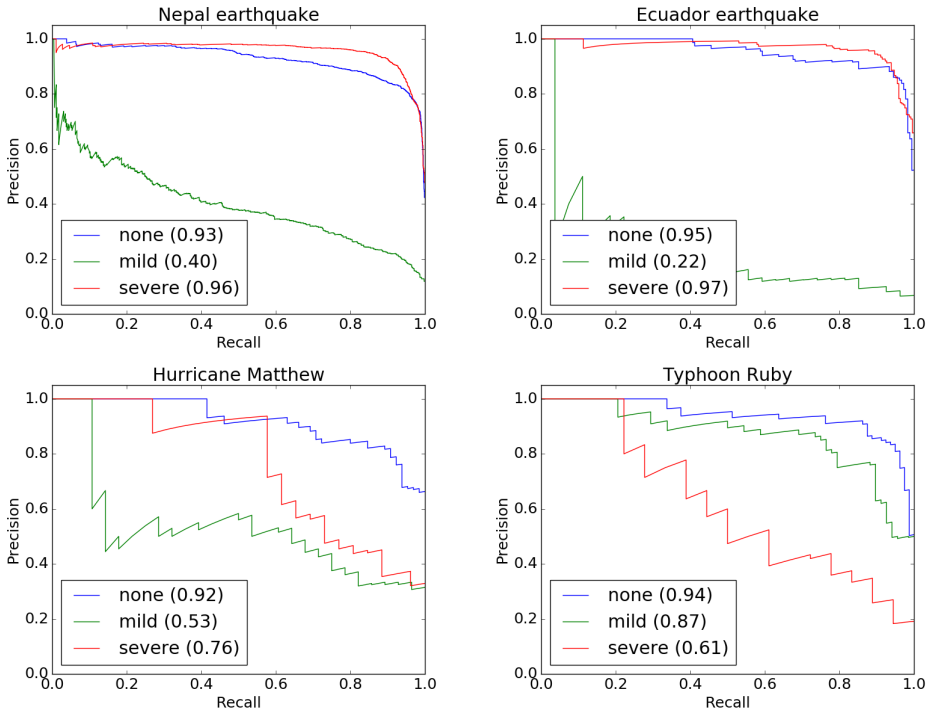

Fig. 2. Precision-recall curves and corresponding AUC scores for all three damage levels (i.e., classes) obtained from the event-specific models.

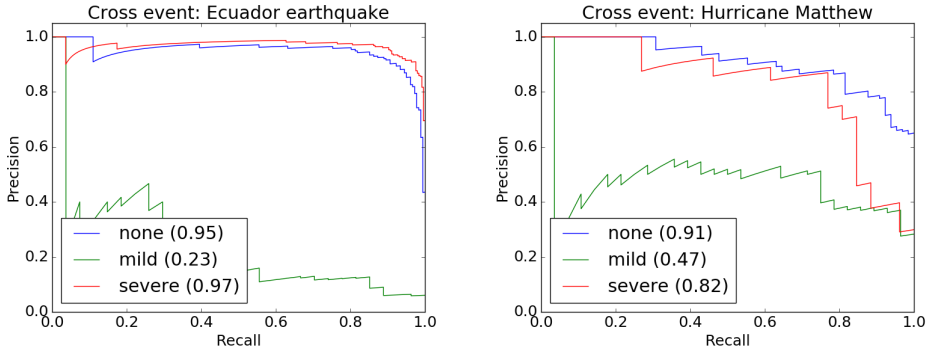

Fig. 3. Precision-recall curves and corresponding AUC scores for all three damage levels (i.e., classes) obtained from the cross-event models.

and recovery operations. Our results show that our approach can be used to classify severity of damage quickly.

Image characteristic differ based on regional conditions, e.g., in some places buildings may be mostly made of bricks, whereas at others they may be made of wood. Consequently, the rubble after the damage in these different regions will look different. Therefore, the best results will always be obtained if we can also get labeled training data from the same event. However, especially at the beginning of a disaster such labeled data may not be readily available. We have shown that using our existing data based on previous events helps address the cold-start problem that occurs during a disaster and shown that we can still achieve reasonable accuracy using the past and the current event data, when available.

In the future, we may consider defining the cases of ambiguity as two different classes. That is, we may have different classes based on the intensity of the damage and on the spread of the damage. For example, one class can correspond to high intensity damage in a small part of the image (1 of 100 buildings, say), high intensity damage over a large area, low intensity damage over a small area and low intensity damage over a large area. Whether this should be done as a multi-class classification or as a hierarchical stacked classification needs to be determined. Also, in the case of stacked classifiers, which feature (intensity/area) should be classified first needs to be carefully designed.

VII. Conclusion

Among other uses of social media, gaining situational awareness during disasters is one of the integral requirements that many humanitarian organizations can fulfill by processing

social media data. One of the situational awareness tasks is to determine the severity of damage to the infrastructure in the disaster zone. For the first time, in this work, we analyzed imagery data posted on social media platforms during disasters for the assessment of the level of damage to infrastructure. Specifically, we used images posted during four natural disasters and contrary to previous research works, which mainly used BoVW models, we proposed and showed the usefulness of deep CNNs with domain-specific fine-tuning to effectively detect the level of damage from images. An extensive set of experiments based on a number of real-world scenarios have been performed on real datasets to show the effectiveness of the proposed approach. We have identified two main challenges, i.e., low prevalence of the training data and non-trivial human-labeling tasks, as our potential future work.

REFERENCES

[1] M. Imran, C. Castillo, F. Diaz, and S. Vieweg, "Processing social media messages in mass emergency: A survey," *ACM Computing Surveys (CSUR)*, vol. 47, no. 4, p. 67, 2015.

[2] C. Castillo, *Big Crisis Data*. Cambridge University Press, 2016.

[3] K. Starbird, L. Palen, A. L. Hughes, and S. Vieweg, "Chatter on the red: what hazards threat reveals about the social life of microblogged information," in *2010 ACM conference on Computer supported cooperative work*, 2010, pp. 241–250.

[4] D. T. Nguyen, K. Al-Mannai, S. R. Joty, H. Sajjad, M. Imran, and P. Mitra, "Robust classification of crisis-related data on social networks using convolutional neural networks." in *ICWSM*, 2017, pp. 632–635.

[5] M. Imran, S. M. Elbassuoni, C. Castillo, F. Diaz, and P. Meier, "Extracting information nuggets from disaster-related messages in social media," *ISCRAM*, 2013.

[6] M. Imran, C. Castillo, J. Lucas, P. Meier, and S. Vieweg, "AIDR: Artificial intelligence for disaster response," in *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 2014, pp. 159–162.

[7] S. Cresci, M. Tesconi, A. Cimino, and F. Dell'Orletta, "A linguistically-driven approach to cross-event damage assessment of natural disasters from social media messages," in *Proceedings of the 24th International Conference on World Wide Web*. ACM, 2015, pp. 1195–1200.

[8] D. T. Nguyen, F. Alam, F. Ofli, and M. Imran, "Automatic image filtering on social networks using deep learning and perceptual hashing during crises," in *ISCRAM*, May 2017.

[9] M. Imran, P. Mitra, and J. Srivastava, "Cross-language domain adaptation for classifying crisis-related short messages," *ISCRAM*, 2016.

[10] M. Everingham, L. Van Gool, C. K. I. 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, Jun. 2010.

[11] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 1097–1105.

[13] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in *International Conference on Learning Representations (ICLR 2014)*.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556

[15] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 1–9.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[17] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *CoRR*, vol. abs/1512.00567, 2015.

[18] S. Zagoruyko and N. Komodakis, "Wide residual networks," *CoRR*, vol. abs/1605.07146, 2016. [Online]. Available: http://arxiv.org/abs/1605.07146

[19] M. D. Zeiler and R. Fergus, *Visualizing and Understanding Convolutional Networks*. Cham: Springer International Publishing, 2014, pp. 818–833.

[20] I. Goodfellow, Y. Bengio, and A. Courville, "Deep learning," 2016, book in preparation for MIT Press. [Online]. Available: http://www.deeplearningbook.org

[21] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[22] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 886–893.

[23] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on Statistical Learning in Computer Vision, ECCV*, 2004, pp. 59–74.

[24] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2006, pp. 2169–2178.

[25] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[26] R. B. Girshick, P. F. Felzenszwalb, and D. McAllester, "Discriminatively trained deformable part models, rel 5," http://people.cs.uchicago.edu/ rbg/latent-release5/.

[27] S. Fidler, R. Mottaghi, A. Yuille, and R. Urtasun, "Bottom-up segmentation for top-down detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013, pp. 3294–3301.

[28] M. Pesaresi, A. Gerhardinger, and F. Haag, "Rapid damage assessment of builtup structures using vhr satellite data in tsunamiaffected areas," *International Journal of Remote Sensing*, vol. 28, no. 13-14, pp. 3013–3036, 2007.

[29] T. Feng, Z. Hong, Q. Fu, S. Ma, X. Jie, H. Wu, C. Jiang, and X. Tong, "Application and prospect of a high-resolution remote sensing and geo-information system in estimating earthquake casualties," *Natural Hazards and Earth System Sciences*, vol. 14, no. 8, pp. 2165–2178, 2014.

[30] M. Turker and B. T. San, "Detection of collapsed buildings caused by the 1999 izmit, turkey earthquake through digital analysis of post-event aerial photographs," *International Journal of Remote Sensing*, vol. 25, no. 21, pp. 4701–4714, 2004.

[31] J. Fernandez Galarreta, N. Kerle, and M. Gerke, "UAV-based urban structural damage assessment using object-based image analysis and semantic reasoning," *Natural Hazards and Earth System Science*, vol. 15, no. 6, pp. 1087–1101, 2015.

[32] F. Ofli, P. Meier, M. Imran, C. Castillo, D. Tuia, N. Rey, J. Briant, P. Millet, F. Reinhard, M. Parkan, and S. Joost, "Combining human computing and machine learning to make sense of big (aerial) data for disaster response," *Big Data*, vol. 4, no. 1, pp. 47–59, Mar 2016.

[33] R. Lagerstrom, Y. Arzhaeva, P. Szul, O. Obst, R. Power, B. Robinson, and T. Bednarz, "Image classification to support emergency situation awareness," *Frontiers in Robotics and AI*, vol. 3, p. 54, 2016.

[34] S. Daly and J. A. Thom, "Mining and classifying image posts on social media to analyse fires," in *ISCRAM*, May 2016, pp. 1–14.

[35] A. Bosch, A. Zisserman, and X. Munoz, "Image classification using random forests and ferns," in *IEEE International Conference on Computer Vision*, 2007.