Crowd-Assisted Disaster Scene Assessment with Human-AI Interactive Attention

Daniel (Yue) Zhang, Yifeng Huang, Yang Zhang, Dong Wang
Department of Computer Science and Engineering, University of Notre Dame
Notre Dame, IN 46556, USA
Email: {yzhang40, yhuang24, yzhang42, dwang5}@nd.edu

Abstract

The recent advances of mobile sensing and artificial intelligence (AI) have brought new revolutions in disaster response applications. One example is *disaster scene assessment (DSA)* which leverages computer vision techniques to assess the level of damage severity of the disaster events from images provided by eyewitnesses on social media. The assessment results are critical in prioritizing the rescue operations of the response teams. While AI algorithms can significantly reduce the detection time and manual labeling cost in such applications, their performance often falls short of the desired accuracy. Our work is motivated by the emergence of crowdsourcing platforms (e.g., Amazon Mechanic Turk, Waze) that provide unprecedented opportunities for acquiring human intelligence for AI applications. In this paper, we develop an interactive Disaster Scene Assessment (iDSA) scheme that allows AI algorithms to directly interact with humans to identify the salient regions of the disaster images in DSA applications. We also develop new incentive designs and active learning techniques to ensure reliable, timely, and cost-efficient responses from the crowdsourcing platforms. Our evaluation results on real-world case studies during Nepal and Ecuador earthquake events demonstrate that iDSA can significantly outperform state-of-the-art baselines in accurately assessing the damage of disaster scenes.

Introduction

Extreme disaster events such as hurricanes and earthquakes can strike a community with little or no warning and leave it with a high level of damages (e.g., casualties, injuries, and infrastructure damages). In such events, the emergency response providers often run out of resources (e.g., rescuers, ambulance, fire trucks) due to the lack of preparation or the sheer volume of the emergency events (Rawls and Turnquist 2010; Zhang et al. 2018a; Wang et al. 2013; 2019; 2012). For example, during the 2017 Hurricane Harvey crisis, 911 emergency lines were overwhelmed with more than 56,000 calls in 15 hours in Houston alone (Gomez 2017). To effectively prioritize the rescue operations, it is critical to accurately assess the damage severity level of the impact areas. Traditionally, such damage assessment tasks are done either manually (e.g., call centers) or through the analysis of remote sensing data (e.g., satellite images). These methods are both expensive and time-consuming. Recently, a new application called Disaster Scene Assessment (DSA) has emerged where deep neural networks (Convolutional Neural Networks (CNNs) in particular) are applied to recognize the severely damaged areas from self-reported social media images of disaster scenes (Li et al. 2018; Nguyen et al. 2017). The automatic tools like CNN allow the DSA algorithms to be much more responsive than manual efforts. The social media images of DSA also provide more detailed on-site information of the disaster from the perspective of the eyewitnesses than the remote sensing data.

Despite the advantages of DSA applications, they suffer from several failure scenarios. For example, they can treat a small fissure on a road in an image as severe damage or take a large collapsed building as moderate damage (Zhang et al. 2019). Such mistakes can mislead the response teams to insignificant events and fail to respond to events where people's lives are at stake. We found one key element that directly contributes to the failure of DSA algorithms is the inaccurate "visual attention" in these algorithms. The visual attention refers to the region of an image that the AI algorithm focuses on to identify the damage level of the scene. We show examples of visual attention (illustrated as heatmap) of three representative DSA algorithms in Figure 1. We also plot the ground truth annotation of human attention (marked in red region) in the figure. Ideally, the visual attention of a DSA scheme should focus on the damaged regions that are the same/similar as those captured by human perception. However, all three DSA schemes in Figure 1 fail to perform such tasks to different extents.

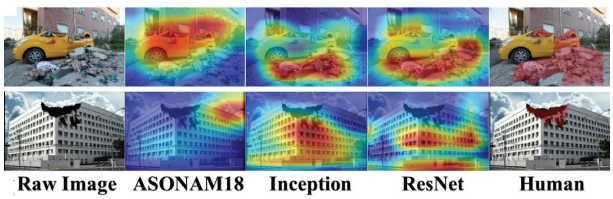


Figure 1: Visual Attention of Various DSA Schemes

A possible approach to improve the visual attention of DSA algorithms is to introduce neural attention mechanisms (Wang et al. 2017) that can emulate the human's visual perception by learning from the labeled data. However, such an approach has several limitations in DSA applications: 1) the lack of dedicated training data for disaster events (which is often caused by expensive labeling costs); 2) the innate algorithm bias caused by the design and structure of the neural network (Lai et al. 2019; Zhang et al. 2019); 3) the noisy nature of the social media images that are taken with diverse camera angles, backgrounds, and resolutions. To address these limitations, we develop a new human-AI DSA system inspired by the observation that the human attention is often much superior and more reliable than the neural attention mechanism in identifying regions of interest in images (Lai et al. 2019). For example, in Figure 1, humans can easily identify the damage areas correctly. A key novelty in this work is to leverage human knowledge to interact with the DSA algorithms to troubleshoot and adjust the attention region of images, which in turn will improve the accuracy of the DSA results. We use the crowdsourcing platforms (e.g., Amazon Mechanic Turk (MTurk)) to obtain human knowledge because they are known for their cost efficiency and the massive amount of freelance workers (Chen, Santos-Neto, and Ripeanu 2012). However, designing such a human-AI interactive system brings some critical technical challenges.

The first challenge lies in the difficulty of understanding and fixing the inaccurate visual attention of AI-based DSA models. We observe that existing attention solutions can be easily misled by a diverse set of objects and the noisy background of disaster scenes. Since these AI-based attention mechanisms are often trained in a black-box fashion, their failure scenarios are hard to explain (Wang et al. 2017) - is it due to the lack of training data? Or, is it due to the wrong design of the attention mechanism? These questions make it non-trivial to leverage the crowd to effectively improve the AI's attention. Current solutions on human-AI systems improves AI primarily by obtaining new ground truth labels and retraining the model (Jarrett et al. 2014; Laws, Scheible, and Schütze 2011). However, such approaches treat an AI model as a black-box and do not intend to understand and troubleshoot its internal attention mechanism. Unfortunately, we found no existing work has been done to leverage the crowd intelligence to troubleshoot and improve AI attention of DSA applications.

The second challenge lies in optimizing the delay-cost trade-off when interacting with the crowdsourcing platform. In particular, DSA applications are often delay-sensitive, requiring all components to respond fast and accurately. In contrast to the AI algorithms where the execution time is quite predictable, the crowd response can be slow and significantly delays the rescue operation. Therefore, it is important to design an incentive mechanism that can effectively stimulate the crowd to provide timely and helpful knowledge to improve the AI algorithm in DSA. However, designing such an incentive mechanism is not a trivial task due to the complex and dynamic relationship between incentives and the response from the crowd.

To address the above challenges, this paper develops a new crowd-AI system - interactive Disaster Scene Assessment (iDSA). To address the first challenge, we propose a new CNN model which designs a novel interactive attention mechanism to allow crowd workers to intervene and adjust the internal visual attention of the DSA model. To our knowledge, iDSA is the first solution to leverage human knowledge from crowdsourcing to directly adjust the internal attention mechanism of AI algorithms in DSA applications. To address the second challenge, we design a constrained multi-arm bandit model to explore the optimal incentive design to acquire timely responses from the crowd workers under strict budget constraints. These designs are integrated into a holistic closed-loop system that allows the AI and crowd to effectively interact with each other and improve the accuracy of both the visual attention and the classification accuracy of DSA. We evaluated the iDSA framework using Amazon MTurk on real-world disaster data traces. Our evaluation results show that iDSA has much better visual attention than the state-of-the-art baselines and consequently achieves a significant accuracy improvement in the DSA applications.

Problem Formulation

In this section, we introduce our AI and crowd models for DSA applications and formally define our problem.

AI-based Disaster Scene Assessment Model

In a DSA application, images posted from social media related to a disaster event are periodically crawled. We refer to each period as a *sensing cycle*. The objective of the DSA application is to classify the damage severity of collected images into different levels such as "no damage", "moderate damage", and "severe damage". We assume that a DSA application has a total of T sensing cycles during a disaster event. For each sensing cycle t, the input data samples to the DSA application is a set of N images, denoted as $X_1^t, X_2^t, ..., X_N^t$, where X_i^t denotes the i^{th} input image at the t^{th} sensing cycle. Each image X_i^t is associated with a ground truth label Y_i^t and an estimated label $\hat{Y}_i^t$ of the damage level. The AI algorithm is pre-trained on a set of training data from previous disaster events and manually labeled by human annotators.

An important aspect of the DSA algorithm is the visual attention. To quantify the accuracy of visual attention, we adopt an Intersection-Over-Union (IOU) metric, which is frequently used to evaluate image segmentation and object detection schemes (Everingham et al. 2010). The IOU metric is defined as $IOU = \frac{Area\ of\ Overlap}{Area\ of\ Union}$, where the "Area of Overlap" and "Area of Union" are computed with respect to a ground truth visual attention, where the damaged area is manually marked by annotators. IOU takes values in [0,1] where 1 represents a complete overlap with the ground truth annotation. We use IOU_i^t to denote the IOU for input X_i^t.

Crowdsourcing Platform Model

We first define the key terms in our crowdsourcing platform.

DEFINITION 1 Crowd Queries ($q(t)$): a set of questions assigned to the crowd at sensing cycle t.

DEFINITION 2 Crowd Responses ($r(t)$): the corresponding answer provided to the crowd query $q(t)$.

We assume each query $q \in q(t)$ is associated with an incentive provided by the application, denoted as b_q. We assume the application has a total budget of B for the crowd. Each response $r \in r(t)$ is associated with a delay denoted as d_r. We observe that the relationship between incentive of a query and the delay of the response from the crowd cannot be simply modeled as linear relationships. Instead, such relationship can be complex and dynamic (Kaufmann, Schulze, and Veit 2011). This observation is critical in the design of the incentive mechanism and quality control schemes in iDSA to ensure timely responses from the crowd.

The goal of iDSA is to maximize the classification accuracy as well as the visual attention accuracy of disaster scenes, while minimizing the average delay from the crowd for a given budget on the crowdsourcing platform. Formally we formulate a constrained multi-objective optimization problem as follows:

$$\begin{aligned} &\text{max: } Pr(Y_i^t = \hat{Y}_i^t | B), \forall 1 \le i \le N, 1 \le t \le T \\ &\text{max: } IOU_i^t, \ \forall 1 \le i \le N, 1 \le t \le T \\ &\text{min: } d_r, \ \forall r \in r(t), 1 \le t \le T \\ &\text{s.t.: } \sum_{t=1}^{T} \sum_{q \in q(t)} b_q \le B \end{aligned} \tag{1}$$

Method

We develop a crowd-assisted interactive Disaster Scene Assessment (iDSA) scheme (Figure 2) to address the problem defined above. The iDSA is designed as a crowd-AI hybrid system that consists of four main modules: i) a Crowd Task Generation (CTG) module; ii) a Budget Constrained Adaptive Incentive (BCAI) module; iii) an Interactive Attention Convolutional Neural Network (IACNN) module; and iv) a Social Media Image Normalization (SMIN) module. We present them in detail below.

Crowd Task Generation (CTG)

The Crowd Task Generation (CTG) module is designed to generate a set of crowd queries to acquire human knowledge to improve the AI-based DSA algorithms. An example query is illustrated in Figure 3. It consists of two parts. The first part is the ground truth annotation where we directly ask the crowd to label their assessment of the damage. The mapping from the annotation score to the class label is discussed in detail in Evaluation. The ground truth annotation allows the AI algorithm to obtain more training data on the fly. The second part of the query is attention annotation where we ask participants to draw the region in the image that they focus on when they assess the damage. The attention annotation allows the AI model to troubleshoot and improve its internal attention mechanism for better performance (discussed in the next subsection). Due to the budget constraint, it is impractical to send all data samples (i.e., images) for the crowd to label (Laws, Scheible, and Schütze 2011). In CTG, we selectively choose images to query the crowd based on two criteria: i) *uncertainty*: the images that cannot be confidently identified by the DSA algorithms should be prioritized in the queries; ii) *diversity*: keeping the annotated data samples diverse will help avoid the repetitive crowd annotations on similar images.

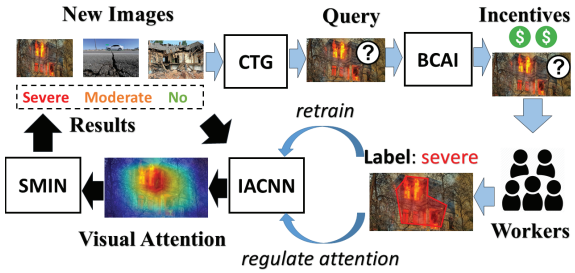


Figure 2: iDSA Overview

Figure 3: Example Query in CTG on MTurk

We design a Query by Committee (QBC)-based active learning (AL) (Seung, Opper, and Sompolinsky 1992) scheme to derive the uncertainty of the DSA algorithms. In particular, we choose a diverse set of M state-of-the-art DSA algorithms $AI_1, AI_2, ..., AI_M$ (referred to as a "committee"). The committee introduces robustness by removing the bias of a single DSA scheme. At a given sensing cycle, each algorithm independently labels all the unseen data samples. We use $Y_{m,i}^t$ to denote the output of AI_m for a given data sample X_i^t. For each algorithm AI_m, we define a weight - w_m^t, representing the authority of the algorithm. We discuss the weight assignment in the IACNN module. The committee decides the classification result of X_i^t as:

$$Y_i^t = \sum_{m=1}^{M} w_m^t \times Y_{m,i}^t \tag{2}$$

Y_i^t is normalized by $\sum_{y \in Y_i^t} y = 1$ where y denotes the probability of a class label in the final output Y_i^t. Then we calculate an entropy score $\mathcal{H}_i^t$ for each DSA algorithm as:

$$\mathcal{H}_i^t = -\sum_{y \in Y_i^t} Pr(y) \times \log Pr(y) \tag{3}$$

We note that highly ranked images in terms of entropy score can be of high similarity (e.g., images all related to road

damages, or images look alike). This is not ideal based on the diversity criterion. Therefore, we design a redundancy filtering algorithm to regulate the diversity of the crowd queries. With a total of K^t queries for the crowd under a given budget, we first assign a pool of K^t candidate images with the highest entropy scores. Then we iteratively remove images from the pool that are significantly similar to others, until all images in the pool have similarity score lower than a predefined threshold. When removing an image, a new image with the next highest entropy is added into the pool. The similarity scores are calculated using the deep auto encoding technique in (Dosovitskiy and Brox 2016).

Budget Constrained Adaptive Incentive Module

After the queries are generated, we design a Budget Constrained Adaptive Incentive (BCAI) module to incentivize the crowdsourcing platform for timely responses to the queries from the crowd. We found that the incentive design problem can be nicely mapped to a constrained multi-armed bandit (CMB) problem in reinforcement learning. The key reason for choosing the bandit solution is that it allows the CMB to dynamically adapt to the uncertain crowdsourcing environments and derive the optimized incentive policy.

We consider a CMB with an action set $\mathcal{A}^t = \{1, 2, ..., Z\}$ at each sensing cycle t, where each entry in $\mathcal{A}$ denotes the amount of money (in cents). We assume each action $z \in \mathcal{A}^t$ generates a non-negative payoff p_z^t (representing the inverse of the crowd response delay) with cost c_z^t at each sensing cycle. The payoff is only revealed at the end of the cycle (i.e., delay is unknown until the responses are submitted by the crowd). We use $\mathcal{C}^t$ to denote the costs from all actions taken at sensing cycle t. The objective of CMB is to derive an optimal incentive policy to maximize the payoffs while keeping the total action cost within the resource budget. The objective is formulated as:

$$\begin{aligned} &\underset{\mathcal{A}^t}{\operatorname{argmax}} \sum_{t=1}^{T} P^t, 1 \le t \le T \text{ (payoff maximization)} \\ &\text{s.t.:} \sum_{t=1}^{T} \mathcal{C}^t \le B, 1 \le t \le T \text{ (budget constraint)} \end{aligned} \tag{4}$$

This objective function can be solved using the classical Epsilon-first policies approach in (Tran-Thanh et al. 2010).

Interactive Attention Convolutional Neural Network Module

Next, we present the Interactive Attention Convolutional Neural Network (IACNN) that leverages an interactive attention design to identify damage areas in disaster scenes. An overview of IACNN is shown in Figure 4. The IACNN employs the deep convolutional neural network, which has been a popular and effective tool for the image classification task (Deng et al. 2009). Our CNN model contains 5 convolutional blocks (with 16 convolutional layers and 5 pooling layers) as shown in Figure 4. We initialize our model with the pre-trained VGG19 model for all convolutional blocks, and fine-tune it using disaster-related images. The existing

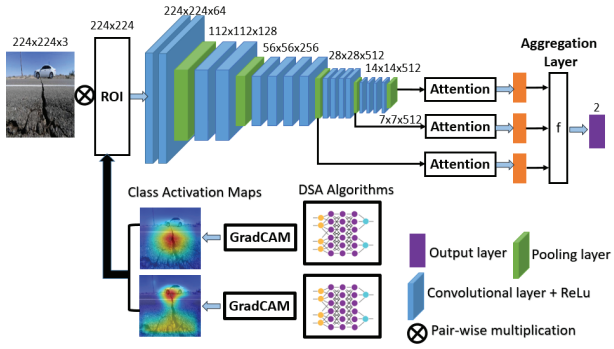


Figure 4: IACNN Overview

DSA algorithms lack explicit attention mechanism to pinpoint the damaged area in the image. To address this issue, our IACNN model develops two attention mechanisms. The first one is a trainable gated attention mechanism that is an internal component of the CNN model. We employ the gated attention approach from (Schlemper et al. 2018), where three separate attention blocks (connected to the last three pooling layers) are aggregated and connected to the final output layer. Compared to existing single attention block approach such as Residual Attention Network (Wang et al. 2017), the gated attention allows the CNN to capture the attention of different resolutions of an image and is more robust against low resolution and noisy image inputs. However, this internal attention alone is not enough to accurately capture the damage region given the limited amount of training data (Lai et al. 2019). Therefore, we design the second interactive external attention mechanism to further enhance the attention of IACNN.

The intuition of the interactive attention is to develop an ensemble of the visual attention from a set of DSA algorithms to decrease the bias of the attention of each individual algorithm. The attention annotations from the crowd are leveraged to derive the weight of each algorithm in the ensemble. In particular, we design an attention ensemble approach by employing the class activation map (CAM) (Selvaraju et al. 2017) technique. The CAM is a visualization technique that can identify the important regions (i.e., pixels) that contribute significantly to the final classification results. Following (Li et al. 2018), we use the last convolutional layer to derive the CAM. Assuming the dimension of the last convolutional layer of ICNN is $U \times V \times L$ (e.g, $14 \times 14 \times 512$ in the proposed CNN), we calculate the CAM score $s_{u,v}$ for each image region (u, v) as:

$$\begin{aligned} s_{u,v} &= \sum_{l \in L} \Big(\lambda_l \times f_l(u, v) \Big) \\ \lambda_l &= \frac{1}{U \times V} \sum_{u \in U, v \in V} \frac{\partial Y}{\partial f_l(u, v)} \end{aligned} \tag{5}$$

where λ_l is a gradient-based weight parameter for the last convolutional layer, and $f_l(u, v)$ represents the value at image location (u, v) in the l-th feature vector. λ_k is derived as the sum of the gradients of output Y with respect to $f_l(u, v)$.

To ensemble the CAMs from each DSA algorithm, we can either 1) find the union of the CAMs; 2) find the intersection of the CAMs; or 3) find the weighted sum of the CAMs. In this work, we pragmatically pick the last approach because it gives the best empirical performance. The weights of each CAM is determined based on how similar it is as compared to the ground truth attention region provided by the crowd workers. We calculate the weight of each DSA algorithm as:

$$w_m^t \propto \sum_{i=1}^{N} IOU_i^t; \quad \sum_{m=1}^{m} w_m^t = 1, \forall 1 \leq t \leq T \quad (6)$$

After combing the CAMs of DSA algorithms, we generate a binary map called Region of Interest (ROI) (Eppel 2017). This binary map is used as a preprocessing layer to the input of IACNN to first filter out the irrelevant regions of image and focus only on the potentially important areas (i.e., damages). The binary map is calculated as:

$$ROI_{u,v} = \begin{cases} 1, & \sum_{m=1}^{M} w_m^t \times s_{u,v,m} > \Theta \\ 0, & \sum_{m=1}^{M} w_m^t \times s_{u,v,m} \leq \Theta \end{cases} \quad (7)$$

where $ROI_{u,v}$ is the ROI score of region (u, v) of the image. $s_{u,v,m}$ is the CAM at region (u, v) from AI_m. Θ is a threshold parameter. We set the Θ to be a relatively small value so it provides a rough filtering of non-important regions based on the CAMs. Then the more fine-grained attention is captured by the trainable gated attention described above.

Social Media Image Normalization Module

Finally, we develop a Social Media Image Normalization (SMIN) module to handle the noisy input of social media images and generate the final classification output. Note that in the IACNN module, we only output two classes - "damage" and "no damage". The reason we chose two classes for IACNN is that the boundary between damage severity levels such as "moderate" and "severe" is unclear to the AI algorithms and they often output the wrong results. Therefore, we adopt the approach from (Doshi, Basu, and Pang 2018) where the damage severity is derived as the percentage of damage regions (e.g., captured by the CAM) in the image. However, the social media images can be taken with diverse camera angles where the absolute size of the damage region cannot always reflect the damage severity levels (Figure 5). To address this issue, we designed a normalized damage score as $\sum_{u \in U, v \in V} \left(\frac{s_{u,v}}{U \times V} \right) * \delta$, where δ is a weighting factor denoting the level of "zoom in" of an image. To calculate δ, we first identify the anchor objects in the images (e.g., cars, bridges, and road signs) using the YOLO V3 object detection tool (Redmon et al. 2016). We then compare the actual size of the anchor objects (based on prior knowledge) with the size of the objects in the image. We observed such normalization significantly improves the classification results as discussed in the evaluation.

Evaluation

In this section, we conduct extensive experiments on real-world datasets to answer the following questions:

(a) Classified as Severe (✗) (b) Classified as Moderate (✓)

Figure 5: Example Failure Scenarios

- **Q1:** Can iDSA achieve a better classification accuracy than the state-of-the-art DSA algorithms?
- **Q2:** Can the interactive attention in iDSA accurately capture the damaged area of a social media image?
- **Q3:** Can iDSA achieve a high crowd responsiveness for DSA applications given a limited budget?
- **Q4:** How does each component of iDSA contribute to its overall performance?

Dataset, Experiment Setup, and Baselines

Data. We use a dataset (Nguyen et al. 2017) that consists of a total of 21,384 social media images related to two disaster events - the 2016 Ecuador Earthquake (2,280 images) and the 2015 Nepal Earthquake (19,104 images) . The dataset contains ground truth labels of damage severity levels. We further collect the ground truth of the exact damaged areas in the image were labeled by multiple human annotators via the LabelMe tool (Russell et al. 2008). We use Amazon MTurk, one of the largest crowdsourcing platforms, to acquire human intelligence for iDSA. In particular, we choose 3 workers to assess the damage severity level of each queried image using the scale from 0 to 5 (see Figure 3). We then use the following rubrics to decide the class label of the image: the aggregated score (from three workers) > 10: severe damage; the aggregated score ≤ 1: no damage; otherwise: moderate damage. For the attention annotation, we treat a pixel in the image as part of the visual attention if it appears in the annotation from at least two workers. For each crowd response, we assign 6 incentive levels (2 cents, 4 cents, 6 cents, 8 cents, 10 cents, and 20 cents) decided by the BCAI module.

Experiment Setup. In our experiments, the dataset is split into a *training set* and a *test set*. The training set contains all 19,104 images from Nepal Earthquake and the test set includes all images from the Equador Earthquake. The choice of data from different events for training and test sets is to emulate the real-world DSA scenarios where the training data is often acquired from the disasters that happened in the past. All compared schemes were run on a server with Intel Xeon E5-2637 v4 3.50GHz CPU and 4 NVIDIA GTX 1080Ti GPUs.

Baselines. We choose a few AI-only algorithms as our DSA baselines, including **ASONAM18** (Li et al. 2018), **Inception** (Szegedy et al. 2016), **ResNet** (He et al. 2016), and **VGGATT** (Wang et al. 2017). We further consider 3 state-of-the-art human-AI hybrid baselines.

- **DirectEnsemble**: It directly ensembles the outputs of the above AI-only schemes (Zhang et al. 2019).
- **Hybrid-Para**: A human-AI hybrid system where the labels from humans and AI algorithms are integrated using a complexity index (Jarrett et al. 2014).
- **Hybrid-AL**: An active learning framework where the annotated labels collected from humans are used to re-train the AI algorithms (Laws, Scheible, and Schütze 2011).

For a fair comparison, we let all AI-only algorithms to randomly query the same amount of images as the human-AI schemes from the crowd. The obtained labels are used as ground truth to retrain the DSA algorithm of the baseline.

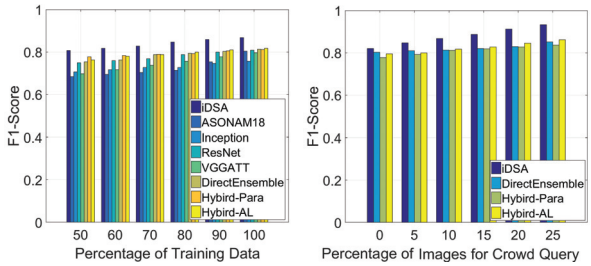
Evaluation Results

Classification Effectiveness (Q1). In the first set of experiments, we focus on the overall performance of all schemes in terms of classification effectiveness, which is evaluated using the classic metrics for multi-class classification: *Accuracy*, *Precision*, *Recall* and *F1-Score*. Similar to (Li et al. 2018; Nguyen et al. 2017), these scores are *macro-averaged* since our dataset has balanced class labels.

Table 1: Classification Accuracy for All Schemes

Algorithms	Accuracy	Precision	Recall	F1
iDSA	**0.869**	**0.881**	**0.853**	**0.867**
ASONAM18	0.808	0.819	0.79	0.804
Inception	0.755	0.75	0.765	0.757
ResNet	0.812	0.819	0.8	0.809
VGGATT	0.799	0.803	0.791	0.797
DirectEnsemble	0.817	0.828	0.799	0.813
Hybrid-Para	0.814	0.826	0.797	0.811
Hybrid-AL	0.821	0.828	0.809	0.818
Gain(%)	5.8	6.4	5.4	6.0

The results are reported in Table 1. We observe iDSA consistently outperforms all baselines. Compared to AI-only schemes, iDSA is able to achieve 7.8% higher F-1 score than the best performing baseline (i.e., ResNet). The reason is that the iDSA can effectively incorporate human intelligence into the DSA algorithm. iDSA is also superior to the human-AI hybrid baselines. For example, iDSA achieved 6% improvement on F-1 score compared to the best-performing human-AI baseline (i.e., Hybrid-AL). This is because none of the hybrid baselines addresses the innate issue of inaccurate attention of the CNN models they use. For example, DirectEnsemble simply aggregates the results of all AI-only algorithms without touching the internal model of these algorithms. In contrast, iDSA directly interacts with the internal attention mechanism of AI and leverages human intelligence to troubleshoot, calibrate, and eventually improve the attention of AI. We further evaluate all schemes by tuning the size of training data (in terms of percentage of the whole training set as shown in Figure 6a) as well as the number of images that we use to query the crowd (from 5% to 25% of the testing set as shown in Figure 6). We observe iDSA outperforms the baselines with different sizes of training data and different amount of knowledge from the crowd.



(a) F1 vs. Training Size (b) F1 vs. Query Size

Figure 6: F1 vs. Training and Query Sizes

Attention Accuracy (Q2). We further investigate whether iDSA can outperform baselines in terms of correctly attending to the damaged areas in the images. We use the IOU metric defined in Problem Formulation. Considering the accuracy of the attention is directly affected by the number of answers collected from the crowd, we tuned the percentage of images that we send out to query the MTurk from 5% to 25%. The results are presented in Table 2. We skip DirectEnsemble and Hybrid-Para in the table because tuning the number of crowd queries will not affect their attention since they do not retrain their models. We observe that iDSA continues to outperform all baselines. The improved accuracy in both attention (Table 2) and classification (Table 1) also validates our hypothesis that improving the attention detection accuracy of DSA algorithms will eventually boost the classification performance of DSA applications.

Table 2: IOU for All Schemes w.r.t % of Images

Algorithms	5% (Images)	10%	15%	20%	25%
iDSA	**0.549**	**0.558**	**0.570**	**0.582**	**0.598**
ASONAM18	0.389	0.392	0.397	0.401	0.405
Inception	0.316	0.322	0.335	0.337	0.341
ResNet	0.451	0.461	0.479	0.490	0.496
VGGATT	0.322	0.329	0.339	0.352	0.358
Hybrid-AL	0.462	0.469	0.474	0.482	0.499
Gain(%)	18.8	19.5	20.2	20.4	19.8

Crowd Responsiveness (Q3). We then evaluate the delay of all human-AI hybrid schemes in terms of 1) execution time, and 2) delay of query answered by the crowdsourcing platform. The results are shown in Table 3. We observe that the response delay from the crowdsourcing platform is the major contributor to the overall delay of human-AI hybrid systems including iDSA. This observation further demonstrates the importance of designing an effective incentive policy to minimize the delay from the crowd and provide timely response to the DSA applications. The results show

that iDSA scheme significantly reduces the crowd delay by 16.8%, 27.3%, and 18.9% compared to DirectEnsemble, Hybrid-Para, and Hybrid-AL, respectively, which all adopt a fixed incentive policy. We attribute such a performance gain to our adaptive incentive module that leverages a multi-arm bandit scheme to dynamically identify the optimal incentive strategy to reduce the response delay from the crowd.

Table 3: Average Delay (in Seconds) per Sensing Cycle

Algorithms	Algorithm Delay	Crowd Delay	Total Delay
iDSA	**66.98**	**427.81**	**494.79**
DirectEnsemble	55.62	513.24	568.86
Hybrid-Para	94.28	588.75	683.03
Hybrid-AL	53.54	527.61	581.15

Abrasion Study (Q4). Finally, we perform a comprehensive *abrasion study* to examine the effect of each component of iDSA. In particular, we present the classification results by removing each of the four modules of iDSA. We found that, by adding the interactive attention design, iDSA is able to increase its F-1 score by 5.2%. The incentive module contributes to 2.8% increase in F-1 score, which highlights the importance of minimizing the response delay from the crowd. We also found the crowd task generation is indeed helpful - yielding 3.5% higher F-1 score. The normalization module helps iDSA to achieve 6.25% performance again by making iDSA more robust against noisy social media images.

Table 4: Abrasion Study

Algorithms	Accuracy	Precision	Recall	F1
iDSA	**0.869**	**0.881**	**0.853**	**0.867**
iDSA w/o CTG	0.839	0.842	0.835	0.838
iDSA w/o IACNN	0.826	0.831	0.818	0.824
iDSA w/o SMIN	0.821	0.842	0.791	0.816
iDSA w/o BCAI	0.843	0.849	0.835	0.842

Related Work

Disaster scene assessment (DSA) is a critical step in disaster response that determines the severity of the damage caused by a disaster based on imagery data. Recently, social media images have become a new data source for DSA which provides more detailed on-site information from the perspective of the eyewitnesses of the disaster at a much lower cost (Rashid et al. 2019). Nguyen *et al.* developed the first deep CNN model with domain-specific fine-tuning to effectively detect the level of damage from social media images (Nguyen et al. 2017). Many follow up solutions have been developed (Li et al. 2018; Zhang et al. 2019; 2018b). However, existing AI-driven solutions are prone to focus on wrong regions of the disaster scene image and provide inaccurate assessment results due to the lack of an explicit attention mechanism.

Recognizing the importance of visual attention in disaster assessment, more recent DSA applications leverage post-hoc attention analysis with the goal of pin-pointing the damaged areas within disaster scene images. For example, Li *et al.* combined CNN and Grad-CAM to generate a heatmap of a given image to locate the damaged area (Li et al. 2018). Recently, trainable attention mechanisms, such as Residual attention (Wang et al. 2017) and SCA-CNN (Chen et al. 2017) have attracted enormous interest in image classification tasks due to their potential in simulating human's visual perception process to improve the classification accuracy. However, these pure data-driven attention mechanisms cannot be directly applied to our problem due to their inferior performance caused by either insufficient training data, or the internal drawbacks of the neural network design. In this work, we propose a novel interactive attention design that leverages human knowledge to troubleshoot and intervene in the attention module in DSA.

A few relevant human-AI hybrid frameworks have been recently developed (Jarrett et al. 2014; Nushi et al. 2017) that allows AI systems to interact with human to improve their performance. Active Learning (AL) is a commonly used technique to combine AI and human intelligence, where AI actively obtains labels of some instances from domain experts (Laws, Scheible, and Schütze 2011). The major benefit of such a framework is that it significantly reduces the labeling costs and improves the efficiency by judiciously selecting a "subset" of data samples to be labeled. However, the AL-based solutions largely ignored the innate limitations of the AI algorithms that cannot be simply improved by retraining the model with more data. In contrast, iDSA is the first solution to leverage human knowledge from crowdsourcing to directly adjust the internal attention mechanism of AI algorithms in DSA applications.

Conclusion

This paper presents iDSA to addresses fundamental challenges in melding crowd intelligence into AI in boosting the performance of DSA applications. iDSA designs a novel interactive attention-aware CNN model to accurately capture the damaged area in the image by interacting with the crowd. An adaptive incentive design is developed to ensure iDSA can acquire timely and reliable responses from the crowd. Evaluation results on a real-world DSA application show that iDSA significantly outperforms existing AI-only DSA solutions and state-of-the-art human-AI frameworks.

Acknowledgement

This research is supported in part by the National Science Foundation under Grant No. CNS-1845639, CNS-1831669, Army Research Office under Grant W911NF-17-1-0409. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; and Chua, T.-S. 2017. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5659–5667.

Chen, X.; Santos-Neto, E.; and Ripeanu, M. 2012. Crowdsourcing for on-street smart parking. In *Proceedings of the second ACM international symposium on Design and analysis of intelligent vehicular networks and applications*, 1–8. ACM.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 248–255. IEEE.

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

Dosovitskiy, A., and Brox, T. 2016. Generating images with perceptual similarity metrics based on deep networks. In *Advances in neural information processing systems*, 658–666.

Eppel, S. 2017. Setting an attention region for convolutional neural networks using region selective features, for recognition of materials within glass vessels. *arXiv preprint arXiv:1708.08711*.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision* 88(2):303–338.

Gomez, L. 2017. Hurricane harvey: 50 counties flooded, 30,000 people in shelters, 56,000 911 calls in just 15 hours. san diego union tribune.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Jarrett, J.; Saleh, I.; Blake, M. B.; Malcolm, R.; Thorpe, S.; and Grandison, T. 2014. Combining human and machine computing elements for analysis via crowdsourcing. In *Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), 2014 International Conference on*, 312–321. IEEE.

Kaufmann, N.; Schulze, T.; and Veit, D. 2011. More than fun and money. worker motivation in crowdsourcing-a study on mechanical turk. In *AMCIS*, volume 11, 1–11. Detroit, Michigan, USA.

Lai, Q.; Wang, W.; Khan, S.; Shen, J.; Sun, H.; and Shao, L. 2019. Human\textit {vs} machine attention in neural networks: A comparative study. *arXiv preprint arXiv:1906.08764*.

Laws, F.; Scheible, C.; and Schütze, H. 2011. Active learning with amazon mechanical turk. In *Proceedings of the conference on empirical methods in natural language processing*, 1546–1556. Association for Computational Linguistics.

Li, X.; Caragea, D.; Zhang, H.; and Imran, M. 2018. Localizing and quantifying damage in social media images. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE.

Nguyen, D. T.; Ofli, F.; Imran, M.; and Mitra, P. 2017. 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*. ACM.

Nushi, B.; Kamar, E.; Horvitz, E.; and Kossmann, D. 2017. On human intellect and machine failures: Troubleshooting integrative machine learning systems. In *AAAI*, 1017–1025.

Rashid, M. T.; Zhang, D.; Liu, Z.; Lin, H.; and Wang, D. 2019. Collabdrone: A collaborative spatiotemporal-aware drone sensing system driven by social sensing signals. In *2019 28th International Conference on Computer Communication and Networks (ICCCN)*, 1–9. IEEE.

Rawls, C. G., and Turnquist, M. A. 2010. Pre-positioning of emergency supplies for disaster response. *Transportation research part B: Methodological* 44(4):521–534.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 779–788.

Russell, B. C.; Torralba, A.; Murphy, K. P.; and Freeman, W. T. 2008. Labelme: a database and web-based tool for image annotation. *International journal of computer vision* 77(1-3):157–173.

Schlemper, J.; Oktay, O.; Chen, L.; Matthew, J.; Knight, C.; Kainz, B.; Glocker, B.; and Rueckert, D. 2018. Attention-gated networks for improving ultrasound scan plane detection. *arXiv preprint arXiv:1804.05338*.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 618–626.

Seung, H. S.; Opper, M.; and Sompolinsky, H. 1992. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, 287–294. ACM.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Tran-Thanh, L.; Chapman, A.; de Cote, E. M.; Rogers, A.; and Jennings, N. R. 2010. Epsilon–first policies for budget–limited multi-armed bandits. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*.

Wang, D.; Kaplan, L.; Le, H.; and Abdelzaher, T. 2012. On truth discovery in social sensing: A maximum likelihood estimation approach. In *Proc. ACM/IEEE 11th Int Information Processing in Sensor Networks (IPSN) Conf*, 233–244.

Wang, D.; Abdelzaher, T.; Kaplan, L.; and Aggarwal, C. C. 2013. Recursive fact-finding: A streaming approach to truth estimation in crowdsourcing applications. In *The 33rd International Conference on Distributed Computing Systems (ICDCS'13)*.

Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; and Tang, X. 2017. Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3156–3164.

Wang, D.; Szymanski, B. K.; Abdelzaher, T.; Ji, H.; and Kaplan, L. 2019. The age of social sensing. *Computer* 52(1):36–45.

Zhang, D. Y.; Badilla, J.; Zhang, Y.; and Wang, D. 2018a. Towards reliable missing truth discovery in online social media sensing applications. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE.

Zhang, Y.; Lu, Y.; Zhang, D.; Shang, L.; and Wang, D. 2018b. Risksens: A multi-view learning approach to identifying risky traffic locations in intelligent transportation systems using social and remote sensing. In *2018 IEEE International Conference on Big Data (Big Data)*, 1544–1553. IEEE.

Zhang, D. Y.; Zhang, Y.; Li, Q.; Plummer, T.; and Wang, D. 2019. Crowdlearn: A crowd-ai hybrid system for deep learning-based damage assessment applications. In *Distributed Computing Systems (ICDCS), 2019 IEEE 39th International Conference on*. IEEE.