Neural Computing and Applications
https://doi.org/10.1007/s00521-022-07717-0

ORIGINAL ARTICLE

MEDIC: a multi-task learning dataset for disaster image classification

Firoj Alam[1] · Tanvirul Alam[2] · Md. Arid Hasan[3,4] · Abul Hasnat[5] · Muhammad Imran[1] · Ferda Ofli[1]

Received: 6 June 2022 / Accepted: 8 August 2022

Abstract

Recent research in disaster informatics demonstrates a practical and important use case of artificial intelligence to save human lives and suffering during natural disasters based on social media contents (text and images). While notable progress has been made using texts, research on exploiting the images remains relatively under-explored. To advance image-based approaches, we propose MEDIC (https://crisisnlp.qcri.org/medic/index.html), which is the largest social media image classification dataset for humanitarian response consisting of 71,198 images to address four different tasks in a multi-task learning setup. This is the first dataset of its kind: social media images, disaster response, and multi-task learning research. An important property of this dataset is its high potential to facilitate research on *multi-task learning*, which recently receives much interest from the machine learning community and has shown remarkable results in terms of memory, inference speed, performance, and generalization capability. Therefore, the proposed dataset is an important resource for advancing image-based disaster management and multi-task machine learning research. We experiment with different deep learning architectures and report promising results, which are above the majority baselines for all tasks. Along with the dataset, we also release all relevant scripts (https://github.com/firojalam/medic).

Keywords Multi-task learning · Social media images · Image classification · Natural disasters · Crisis informatics · Deep learning · Dataset

1 Introduction

Natural disasters cause significant damage (e.g., Hurricane Harvey in 2017 cost $125 billion)[1] and require urgent assistance in time of crisis. In the last decade, various social media played important roles in humanitarian response tasks as they were widely used to disseminate information and obtain valuable insights. During disaster events, people post content (e.g., text, images, and video) on social media to ask for help (e.g., report of a person stuck on a rooftop during a flood), offer support, identify urgent needs, or share their feelings. Such information is helpful for humanitarian organizations to take immediate actions to plan and launch relief operations. Recent studies demonstrated that images shared on social media during a disaster can assist humanitarian organizations in recognizing damages in infrastructure [1], assessing damage severity [2], identifying humanitarian information [3], detecting crisis incidents [4], and detecting disaster events with other related tasks [5]. However, the amount of

Available at: https://crisisnlp.qcri.org/medic/index.html
https://github.com/firojalam/medic

✉ Firoj Alam
fialam@hbku.edu.qa

Tanvirul Alam
tanvirul.alam@mail.rit.edu

Md. Arid Hasan
arid.cse0325.c@diu.edu.bd

Abul Hasnat
mhasnat@gmail.com

Muhammad Imran
mimran@hbku.edu.qa

Ferda Ofli
fofli@hbku.edu.qa

[1] Qatar Computing Research Institute, HBKU, Doha, Qatar

[2] Rochester Institute of Technology, Rochester, Rochester, USA

[3] Cognitive Insight Limited, Dhaka, Bangladesh

[4] Daffodil International University, Dhaka, Dhaka, Bangladesh

[5] BLACKBIRD.AI, Rochester, USA

[1] https://en.wikipedia.org/wiki/List_of_disasters_by_cost.

Published online: 03 September 2022

research and resources to develop powerful computer vision-based predictive models remains insufficient compared to the NLP-based progress [6–8]. Motivated by these observations, this research aims to enrich available resources to make further advancements in the computer vision-based disaster management studies.

Recent advances in deep convolutional neural networks (CNN) and their learning techniques provide efficient solutions for different computer vision applications. While simple applications can be realized with a single-task formulation such as classification [9], semantic segmentation [10], or object detection [11], the complex ones such as autonomous vehicles, robotics, and social media image analysis [12, 13] necessitate incorporating multiple tasks, which significantly increases the computational and memory requirements for both training and inference. Multi-task learning (MTL) techniques [13–15] have emerged as the standard approach for these complex applications where a model is trained to solve multiple tasks simultaneously, which helps to improve the performance, reduce inference time and computational complexities. For example, an image posted on social media during a disaster event may contain information whether it is a flood event, shows infrastructure damage, and is severe. Such a multitude of information needs to be detected in real-time to help humanitarian organizations [12, 16] with various tasks including (i) disaster type recognition, (ii) informativeness classification, (iii) humanitarian categorization, and (iv) damage severity assessment (see Sect. 3 for more details). Existing works [1–3] present separate task-specific models, resulting in higher computational complexities (e.g., computational power, training and inference time). Hence, this research aims at reducing this overhead by addressing different tasks simultaneously with an MTL setup, which can also help reduce the carbon footprint [17].

Labeled public image datasets, such as ImageNet [18] and Microsoft COCO [19] made significant contributions to the advancement of today's powerful machine learning models. Likewise, for the MTL setup, several image datasets have already been proposed, which are summarized in Table 1. These datasets include images from different domains such as indoor scenes, driving, faces, handwritten digits, and animal recognition, which are already contributing to the advancement of MTL research. However, an MTL dataset for critical real-world applications which comprise humanitarian response tasks during natural disasters is yet to become available. This paper proposes a novel MTL dataset for disaster image classification.

To this end, we build upon the previous work of Alam et al. [5] where the images are mostly annotated for individual tasks, and only 5558 out of 71,198 images have labels for all four tasks mentioned above. We provide an expansive extension by annotating the images for all tasks, i.e., we annotated 155,899 more labels for these tasks in addition to the existing ones.[2] For *disaster type recognition* and *humanitarian categorization* tasks, we also labeled a part of the images with multiple labels following a weak supervision approach as they are suitable for multilabel annotation (see Sect. 3). Figure 1 shows example images with the labels for all four tasks.

Our contributions in this research can be summarized as follows: (i) we provide a social media MTL image dataset for disaster response tasks with various complexities, which can be used as an evaluation benchmark for computer vision research; (ii) we ensured high quality annotations by making sure that at least two annotators agree on a label; (iii) we provide a benchmark for heterogeneous multi-task learning and baseline studies to facilitate future study; (iv) our experimental results can also be used as a baseline in the single-task learning setting.

The rest of the paper is organized as follows. Section 2 provides an overview of the existing work. Section 3 introduces the tasks and describes the dataset development process. Section 4 explains the experiments and presents the results while Sect. 5 provides a discussion. Finally, we conclude the paper in Sect. 6.

2 Related work

This paper mainly focuses on the development of an MTL dataset for disaster response tasks. Therefore, we first review the recent work on MTL and available MTL datasets; and then, survey social media image classification literature and datasets for disaster response.

2.1 Multi-task learning and datasets

Multi-task learning (MTL) aims to improve generalization capability by leveraging information in the training data consisting of multiple related tasks [14]. It simultaneously learns multiple tasks and has shown promising results in terms of generalization, computation, memory footprint, performance, and inference time by jointly learning through a shared representation [14, 15]. Since the seminal work by Caruana [14], MTL research has received wide attention in the last several years in NLP, computer vision, and other research areas [15, 20–23]. MTL brings benefits when associated tasks share complementary information. However, performance can suffer when multiple tasks have conflicting needs, and the tasks have competing priorities (i.e., one is superior to the other). This phenomenon is referred to as negative transfer. This understanding led to

[2] For four tasks, 71,198 images now have 284,792 labels whereas previous annotations comprised only 128,893 labels.

Table 1 Upper part of the table presents the datasets used in multi-task learning studies in computer vision research

Refs.	Dataset	Source	Size	Task type	# Tasks	Tasks	# Classes	Domain	Year
Datasets used for multi-task learning									
[15]	PASCAL [42, 43]	Flickr	12,030 (I)	Hete.	5	SS, HS, SE, and SD	–	Diverse objects	2021
[15]	NYU-V2 [41]	PC	1449 (I)	Hete.	3	IS, SS, and SC	–	Indoor video	2021
[13]	BDD100K	Nexar	100,000 (V)	Hete.	10	Ten tasks	[a]	Driving	2020
[24]	MNIST [34]	–	70,000 (I)	Homo.	10	10 digits cls.	10 CL.	Handwritten	2019
[24]	CIFAR10 [45]	–	60,000 (I)	Homo.	10	10 animal cls.	10 CL.	Animal	2019
[24]	UCSD-Birds [46]	–	11,788 (I)	Homo.	10	10 R/tasks	Ranking	Animal	2019
[24]	OmniGlot [47]	–	1,623 (I)	Homo.	50	50 alphabets	50 CL.	Handwritten	2019
[24]	OmniArt [40]		133,000 (S)	Hete.	7	7 tasks	–	Artwork	2019
[44]	Taskonomy	IC	4M (I)	Hete.	26	26 tasks	–	Indoor scenes	2018
[48]	Office-caltech [36]		2,533 (I)	Homo.	4	Amazon, Webcam, and DSLR, Caltech-256	10 CL/task		2017
[48]	Office-Home [49]	SE	15,500 (I)	Homo.	4	Artistic, clip art, product, and real-world images	65 objects	Office/ Home	2017
[48]	[b]	–	2400 (I)	Homo.	4	Caltech-256, ImageNet Pascal and Bing	–	Diverse	2017
[37]	MNIST [34]	–	70,000 (I)	Homo.	10	10 digits cls.	10 CL.	Handwritten	2016
[37]	AdienceFaces [39]	Flickr	16,252 (I) (G), 16,139 (I) (A)	Hete.	2	Gender, Age	Gender: 2 Age: 8	Face	2016
[33]	USPS [35]	–	2000 (S)	Homo.	10	10 ways/tasks	Digits: 0–9	Handwritten	2012
[33]	MNIST [34]	–	2000 (S)	Homo.	10	10 ways/tasks	Digits: 0–9	Handwritten	2012
[32]	USPS [35]	–	2000 (S)	Homo.	10	10 ways/tasks	Digits: 0–9	Handwritten	2011
[32]	MNIST [34]	–	2000 (S)	Homo.	10	10 ways/tasks	Digits: 0–9	Handwritten	2011
[32]	Animal [50]	–	30,000 (I)	Homo.	20	20 ways/tasks	20 CL.	Animal	2011
Disaster-related datasets									
[4]	Incident	Web, SM	446,684 (I)	NA	1	Incident	43	Incidents	2020
[5]	CrisisBench.	Web, SM	DT:17,511, Info:59,717, Hum:17,769, DS:34,896	NA	4	DT, Info, Hum, DS	DT: 7, Info: 2, Hum:4, DS:3	Disaster	2020
[51]	xBD	Satellite	700,000	NA	–	Building damage	4	Disaster	2019
[52]	MediaEval 2018	SM	1,654 I/P	NA	1	Flood	R. and cls.: 2 CL.	Disaster	2018
[3]	CrisisMMD	SM	18,082	NA	3	Info, Hum, DS	Info: 2, Hum:8, DS:3	Disaster	2018
[1]	DMD	Web	5878	NA	1	Damage	6	Disaster	2018
[53]	DAD	SM	~25,000	NA	1	DS	3	Disaster	2017
[54]	DIRSM	Flickr	T1: 6600 (I); T2: 462 I/P	NA	1	Flood	R, cls.: 2 CL	Disaster	2017
Proposed disaster-related multi-task learning dataset									
	MEDIC	SM	71,198 (I)	Hete.	4	DT, Info, Hum, DS	DT: 7, Info: 2, Hum:4, DS:3	Disaster	2021

Middle part shows disaster related datasets, and the last row shows our proposed dataset. I: Images, V: Videos, S: Samples, SE: Search engines, SM: Social media, DT: disaster types, Info: Informativeness, Hum: Humanitarian, DS: Damage severity. CL.: number of class labels. Hete.: heterogeneous, Homo: Homogeneous. PC: Personal collection. SS: semantic segmentation, HS: human part segmentation, SE: semantic edge detection of surface normals prediction, SD: saliency detection, IS: instance segmentation, SC: scene classification, IC: Indoor scenes, cls.: classification, R/tasks: Ranking tasks, I/P: image patches

[a] http://https://doc.bdd100k.com/format.html#categories

[b] ImageCLEFhttp://http://imageclef.org/2014/adaptation

Fig. 1 Examples of images representing all tasks. **T1**: Disaster types, **T2**: Informativeness, **T3**: Humanitarian, **T4**: Damage severity

the question of what, when, and how to share information among tasks [15, 24]. To address these aspects, in the deep learning era, numerous architectures and optimization methods have been proposed. The architectures are categorized into hard and soft parameter sharing. Hard parameter sharing design consists of a shared network followed by task-specific heads [25–27]. In soft parameter sharing, each task has its own set of parameters, and a feature sharing mechanism to deal with cross-task talk [28–30]. In MTL literature, a problem can be formulated in two different ways—homogeneous and heterogeneous [24]. While the homogeneous MTL assumes that each task corresponds to a single output, the heterogeneous MTL assumes each task corresponds to a unique set of output labels [14, 31]. The latter setting uses a neural network using multiple sets of outputs and losses. In this study, we aim to provide a benchmark with our heterogeneous MTL dataset using the hard parameter sharing approach.

Earlier studies such as [32] and [33] mostly exploited the MNIST [34] and USPS [35] datasets for MTL experiments. These datasets were originally designed for single-task classification settings. For example, the widely used MNIST dataset was originally designed for digit classification, and Office-Caltech [36] was designed to categorize images in 31 classes, which are collected from different domains. However, such datasets are used with the homogeneous problem setting of multi-task learning by selecting ten target classes as ten binary classification tasks [24, 33, 37]. Numerous other widely used datasets such as MC-COCO [19] and CelebA [38] have also been used for multi-task learning in the homogeneous problem setting.

Several existing datasets consisting of multiple unique output label sets were studied in the heterogeneous setting. For example, AdienceFaces [39] was designed for gender and age group classification tasks, OmniArt [40] consists of seven tasks, NYU-V2 [41] consists of three tasks, and PASCAL [42, 43] consists of five tasks. Very few datasets were specifically designed for multi-task learning research. Most notable ones are Taskonomy [44] and BDD100K [13]. The Taskonomy dataset consists of four million images of indoor scenes from 600 buildings, and each image was annotated for twenty-six visual tasks. Ground truths of this dataset were obtained programmatically, and knowledge distillation approaches. The BDD100K dataset is a diverse 100K driving video dataset consisting of ten tasks. It was collected from Nexar,[3] where videos are uploaded by the drivers. In Table 1, we provide widely used datasets, which have been used for MTL.

2.2 Disaster response studies and datasets

During disaster events, social media content has proven to be effective in facilitating different stakeholders including humanitarian organizations [55]. Alongside, there has been growing research interest in developing computational methods and systems to better analyze and extract actionable information from social media content [7, 56, 57]. Most of such efforts relied on social media content, such as Twitter and Facebook, for humanitarian aid [58, 59]. Given that accessing Facebook data became difficult, the use of

[3] https://www.getnexar.com/.

Twitter content remained more popular. Research studies and resource development have focused on Twitter content due to its instant access to timely multi-modal information (i.e., textual and visual) as such information is crucial for different stakeholders (e.g., governmental and non-governmental organizations) [58, 59]. Notable resources with textual content include the CrisisLex [60], CrisisNLP [61], TREC Incident Streams [62], disaster tweet corpus [63], Arabic Tweet Corpus [64], CrisisBench [65], HumAID [66], and CrisisMMD (text and image) [3, 67]. In the past years, several systems have also been developed and deployed during disaster events [58, 68–70]. One notable system is AIDR [58][4], which has been used during major disaster events to collect and classify tweets, and provide a visual summary.

Earlier research efforts in crisis informatics are mainly focused on textual content analysis [8]. However, lately there has been a growing interest on the imagery content analysis as images posted on social media during disasters can play significant role as reported in many studies [2, 12, 16, 53, 71, 72]. Recent works include categorizing the severity of damage into discrete levels [2, 16, 53] or quantifying the damage severity as a continuous-valued index [73, 74]. Such models were also used in real-time disaster response scenarios by engaging with emergency responders [70]. Other related work includes adversarial networks for data scarcity issues [75, 76]; disaster image retrieval [77]; image classification in the context of bush fire emergency [78]; flood photo screening system [79]; sentiment analysis from disaster image [80]; monitoring natural disasters using satellite images [7, 81]; and flood detection using visual features [82].

Publicly available image datasets include damage severity assessment dataset (DAD) [2], multimodal dataset (CrisisMMD) [3] and damage identification multimodal dataset (DMD) [1]. The first dataset is only annotated for images, whereas the last two are annotated for both text and images. Other relevant datasets are Disaster Image Retrieval from Social Media (DIRSM) [54] and MediaEval 2018 [52]. The dataset reported in [51] was constructed for detecting damage as an anomaly using pre-and post-disaster images. It consists of 700,000 building annotations. A similar and relevant work is the Incidents dataset [4], which consists of 446,684 manually labeled images with 43 incident categories. The *Crisis Benchmark Dataset* reported in [5] is the largest social media disaster image classification dataset, which is a consolidated version of DAD, CrisisMMD, DMD, and additional labeled images.

In this study, we extended the *Crisis Benchmark Dataset* to adapt it to an MTL setup. To that end, we assigned images with 155,899 more labels to ensure that the entire dataset contains aligned labels for all the tasks. Additionally, we annotated some images with multiple labels, when appropriate, for humanitarian categorization and disaster type recognition tasks.

3 MEDIC dataset

The MEDIC dataset consists of four different disaster-related tasks that are important for humanitarian aid.[5] These tasks are defined based on prior work experience with the humanitarian response organizations such as UN-OCHA and existing literature [3, 6, 12, 58]. In this section, we first provide the details of each task and class labels, and then, discuss the annotation details of the dataset.

3.1 Tasks

Disaster types

During man-made and natural disasters, people post textual and visual content about the current situation, and the real-time social media monitoring system requires to detect an event when ingesting images from unfiltered social media streams. For the disaster scenario, it is important to automatically recognize different disaster types from the crawled social media images. For instance, an image can depict a wildfire, flood, earthquake, hurricane, and other types of disasters. Different categories (i.e., natural, human-induced, and hybrid) and sub-categories of disaster types have been defined in the literature [83]. This research focuses on major disaster events that include (i) earthquake, (ii) fire, (iii) flood, (iv) hurricane, (v) landslide, (vi) other disaster, which covers all other types (e.g., plane, train crash), and (vii) not disaster, which includes the images that do not show any identifiable disasters.

Informativeness

Social media contents are often noisy and contain numerous irrelevant images such as cartoons and advertisements. In addition to this, the clean images that show damaged infrastructure due to flood, fire, or any other disaster events are crucial for humanitarian response tasks. Therefore, it is necessary to eliminate any irrelevant or redundant content to facilitate crisis responders' efforts more effectively. For this purpose, we define the *informativeness* task as to filter out irrelevant images, where the class labels comprise (i) informative and (ii) not informative.

Humanitarian

Fine-grained categorization of certain information significantly helps the emergency crisis responders to make an efficient actionable decision. Humanitarian categories vary

[4] http://aidr.qcri.org/.

[5] https://en.wikipedia.org/wiki/Humanitarian_aid.

Table 2 Data collection source, event name, year of the event and number of image annotated

Source	Event name	Year	# Images	Source	Event name	Year	# Images
Twitter	Typhoon ruby/hagupit	2014	833	Twitter	Iraq iran earthquake	2017	596
Twitter	Nepal earthquake	2015	21,710	Twitter	Mexico earthquake	2017	1378
Twitter	South India floods	2015	1476	Twitter	Srilanka floods	2017	1022
Twitter	Illapel earthquake	2015	403	Twitter	Ukraine conflict	2017	240
Twitter	Food insecurity in yemen	2015	466	Twitter	Greece wildfire	2018	351
Twitter	Paris attack	2015	1043	Twitter	Hurricane florence	2018	186
Twitter	South India floods	2015	753	Twitter	Hurricane michael	2018	219
Twitter	Syria attacks	2015	350	Twitter	Kerala flood	2018	605
Twitter	Terremotoitalia	2015	919	Twitter	Typhoon mangkhut	2018	172
Twitter	Ecuador earthquake	2016	2280	Google	NA	NA	3007
Twitter	Hurricane matthew	2016	596	Twitter	Human induced disaster	NA	501
Twitter	California wildfires	2017	1585	G, B, F	NA	NA	1263
Twitter	Hurricane harvey	2017	5644	Twitter	Natural disaster	NA	6597
Twitter	Hurricane irma	2017	4,973	Twitter	Security incidents activities	NA	1082
Twitter	Hurricane maria	2017	5069	G, I.	NA	NA	5,879

G: Google, B: Bing, F: Flickr, I: Instagram

depending on the type of content (text vs. image). For example, the CrisisBench dataset [65] consists of tweets labeled with 11 categories, whereas CrisisMMD [3] multimodal dataset consists of eight categories. Such variation exists between text and images because some information can easily be presented in one modality than another modality. For example, it is possible to report *missing or found people* in text than in an image, which is also reported in [3]. This research focuses on these factors and considers the four most important categories that are useful for crisis responders such as (i) affected, injured, or dead people, (ii) infrastructure and utility damage, (iii) rescue volunteering or donation effort, and (iv) not humanitarian.

Damage severity

Detecting the severity of the damage is significantly important to help the affected community during disaster events. The severity of the damage can be assessed from an image based on the visual appearance of the physical destruction of a built structure (e.g., bridges, roads, buildings, burned houses, and forests). In-line with [2], this research defines the following categories for the classification task: (i) severe damage, (ii) mild damage, and (iii) little or none.

3.2 Annotations

3.2.1 Data curation

This research extends the labels of the Crisis Benchmark dataset [5]. The Crisis Benchmark dataset was developed by consolidating existing datasets and labeling new data for disaster types. The Crisis Benchmark dataset consists of images collected from Twitter, Google, Bing, Flickr, and Instagram. The majority of the datasets have been collected from Twitter, as shown in Table 2. The Twitter data were mainly collected during major disaster events[6] and using different disaster-specific keywords. The data collected from Google, Bing, Flickr, and Instagram are based on specific keywords. The dataset is diverse in terms of (i) number of events, (ii) different time frames spanning over five years, (iii) natural (e.g., earthquake, fire, floods) and man-made disasters (e.g., Paris attack, Syria attacks), and (iv) events occurred in different parts of the world. The number of images in different events resulted from different factors, such as the number of tweets collected during the disaster events, the number of images crawled, filtered due to duplicates, and a random selection for the annotation. Our motivation for choosing and extending the Crisis Benchmark dataset is that it reduced the overall cost of data collection and annotation processes while also having a large dataset for MTL.

3.2.2 Multiclass annotation

For the manual annotation, we used Appen[7] crowdsourcing annotation platform. In such a platform, finding qualified workers and managing the quality of the annotation is an important issue. To ensure the quality, we used the widely used gold standard evaluation approach [84]. We designed the interface with annotation guidelines on Appen for the annotation task (see Fig. 7 in Appendix). We followed the

[6] Event names reported in Table 2 are based on Wikipedia.

[7] https://appen.com/.

annotation guidelines from previous work [3, 5] and improved with examples for this task (see the detailed annotation guidelines with examples in Appendix A).

For all tasks, we first annotated images with a multiclass setting. Then for *humanitarian* and *disaster type* tasks we labeled the images with multiple labels as they are more suitable to be framed as pure multilabel setting (see Sect. 3.2.4). For the multiclass labeling, our decision has been influenced by several factors. The most important one was our consultation with humanitarian organizations which suggested limiting the number of classes by merging related ones and keeping only the most important information types. This is due to the information overload issue that humanitarian responders often deal with at the onset of a disaster situation if exposed to information types not important for them. For an image that can have multiple labels, we instructed the annotators to select the label that is more important for humanitarian organizations and prominent in the image.

For the annotation, we designed a *HIT* containing five images. For the gold standard evaluation, we manually labeled 100 images, which are randomly assigned to the HIT for the evaluation. We assigned a criterion to have at least three annotations per image and per task. An agreement score of 66% is used to select the final label, which ensured that at least two annotators agreed on a label. The HIT was extended to more annotators if such a criterion was not met.

Since the Crisis Benchmark dataset did have task-specific labels for all images, i.e., different sets of images consisted of labels for three tasks and two tasks, we first prepared different sets with missing labels for the annotation. For example, 25,731 images of the Crisis Benchmark dataset did not have labels for disaster types and humanitarian tasks, which we selected for the annotation tasks. In this way, we run the annotation tasks in different batches.

3.2.3 Crowdsourcing results

To measure the quality of the annotation, we compute the annotation agreement using Fleiss kappa [85], Krippendorff's alpha [86] and average observed agreement [85]. In Table 3, we present the annotation agreement for all events with different approaches mentioned above. The agreement score varies from 46 to 71% for different tasks. Note that, in the Kappa measurement, the values of ranges 0.41–0.60, 0.61–0.80, and 0.81–1 refers to moderate, substantial, and perfect agreement, respectively [87]. Based on these measurements, we conclude that our annotation agreement score leads to moderate to substantial agreement. The number of labels and subjectivity of the annotation tasks reflected the annotation agreement score. Some annotation tasks are highly subjective. For example, for the disaster-type task, hurricane or tropical cyclones often leads to heavy rain, which causes flood (e.g., an image showing a fallen tree with flood water) can be annotated as hurricane or flood. Another example is an image showing building damage and rescue effort. In such cases, the annotation task was to carefully check what is more visible in the image and select the label accordingly. Note that the agreement score for disaster types is comparatively lower than other tasks, which is due to the high level of subjectivity in the annotation task. Annotators needed to choose one label among seven labels. The average agreement scores are comparatively higher as we made sure at least two annotators agree on a label.

Table 3 Annotation agreement for different tasks

Tasks	Fleiss (κ)	Krip. (α)	Avg agg.
Disaster types	0.46	0.46	0.70
Informativeness	0.71	0.71	0.91
Humanitarian	0.52	0.52	0.73
Damage severity	0.55	0.55	0.79

Fleiss Kappa (κ), Krip. (α): Krippendorff's α, Avg agg.: Average observed agreement

3.2.4 Multilabel annotation

For the multilabel annotation for *disaster types* and *humanitarian* tasks, we followed a weak supervision approach to assign multiple labels due to the annotation budget (e.g., time, cost). We selected and assigned a *set* of labels from all annotators. Given that we have three annotators A_1, A_2, and A_3, who assigned a label l from $\mathbb{L} = \{l_1, l_2 ... l_n\}$ to an image $\mathbb{I}$, the final label set for the image $\mathbb{I}$ is defined as $\mathbb{I}_{\mathbb{L}} = \mathbb{S}\{A_1^l, A_1^l, A_1^l\}$. Here, the label with majority agreement ($\geqslant 66\%$) is the same label as in our multiclass setting, and the rest of the labels can have a lower agreement. Note that we were able to assign multiple labels on 53,683 images (75.4%) for disaster types and 65,038 (91.3%) for humanitarian tasks out of 71,198 images (see Table 5). As images have been labeled in different phases and curated from existing sources, we could not properly manage to have multiple labels for all images.

3.2.5 Resulting dataset

After completing the annotation task, the proposed dataset added 155,899 labels for four tasks in addition to the existing 128,893 labels from 71,198 images. In total, this research re-annotated 65,640 images to create the MEDIC dataset. Furthermore, we enriched the MEDIC dataset by separately providing multilabel annotations for *disaster types* and *humanitarian* tasks. The distributions for

Table 4 Annotated dataset with data splits for different tasks

Label	Train	Dev	Test	Total
Disaster Types				
Earthquake	12,296	1004	1795	15,095
Fire	1796	262	690	2748
Flood	3401	587	1315	5303
Hurricane	4517	651	1518	6686
Landslide	1065	168	331	1564
Not disaster	24,459	3141	8885	36,485
Other disaster	1819	344	1154	3317
Total	49,353	6157	15,688	71,198
Informativeness				
Informative	28,073	3478	7206	38,757
Not informative	21,280	2679	8,482	32,441
Total	49,353	6157	156,88	71,198
Humanitarian				
Affected injured or dead people	3662	274	639	4575
Infrastructure and utility damage	18,994	2440	5224	26,658
Not humanitarian	24,427	3099	9145	36,671
Rescue volunteering or donation effort	2270	344	680	3294
Total	49,353	6157	15,688	71,198
Damage Severity				
Little or none	28,314	3613	10,252	42,179
Mild	3904	698	1527	6129
Severe	17,135	1846	3909	22,890
Total	**49,353**	**6157**	**15,688**	**71,198**

Table 5 Multilabel annotated dataset with data splits for different tasks

Disaster Types					Humanitarian			
# Labels	Train	Dev	Test	Total	Train	Dev	Test	Total
1	32,227	3610	9635	45,472	40,885	4777	11,749	57,411
2	5553	662	1202	7417	5550	491	1019	7060
3	579	88	133	800	445	37	85	567
Total	**38,359**	**4360**	**10,970**	**53,689**	**46,880**	**5305**	**12,853**	**65,038**

multiclass and multilabel annotations are shown in Tables 4 and 5, respectively. We have analyzed the dataset to understand how tasks and the labels are associated with each other, for which we have computed confusion matrices between pairs of tasks. We find a good correlation between labels across tasks. For example, between humanitarian and damage severity tasks, majority of the *not-humanitarian* images are also labeled as *little or none* as shown in Fig. 8d in Appendix A.5. We have similar observations for other task pairs as well. As for the multilabel annotation, majority of the images are labeled with single label. For example, for disaster types 84.7% images are labeled with single label and 15.3% with 2-3 labels. For humanitarian, 88.3% are with single label and rest are 11.7%.

3.3 Comparison with other datasets

A comparative analysis with prior disaster-related datasets suggests that the MEDIC dataset is larger in size, covering aligned labels for four tasks, and containing multilabel annotations. In Table 6, we present a comparison of the datasets containing aligned labels for MTL. From the table, it is clear that the prior datasets are not designed for this kind of learning setup and the distribution of the class labels is highly skewed (see Table 9 in [88] for Crisis Benchmark Dataset).

Table 6 Multi-task learning datasets for disaster image classification tasks

	DT	Info	Hum	DS	Multilabel	Total
CrisisMMD [3]		✓	✓	✓		3533
Crisis Benchmark Dataset [5]	✓	✓	✓	✓		5558
MEDIC	✓	✓	✓	✓	✓	**71,198**

DT: Disaster types, Info: Informativeness, Hum: Humanitarian, DS: Damage severity

4 Experiments and results

In Table 4, we present the dataset with task-wise data splits and distribution for multiclass setting. The distribution for multiclass setting consists of 69%, 9%, and 22% for training, development, and test sets, respectively. We first conduct a baseline experiment, followed by a single-task learning experiment to compare and provide a benchmark for a multi-task setting.

To measure the performance of each classifier and for each task setting, we use weighted average precision (P), recall (R), and F1-score (F1), which are widely used in the literature. For the multilabel experiments we computed micro average precision (P), recall (R), F1-score (F1) and humming loss, which are commonly used metrics [89, 90].

4.1 Baseline

For the baseline experiment we evaluate (i) a majority class baseline, and (ii) fixed features from a pre-trained model used for training and testing SVM and KNN. We extracted features from the penultimate layer of the EfficientNet (b1) model [91], which is trained using ImageNet. The majority class baseline predicts the label based on the most frequent label in the training set. This has been most commonly used in shared tasks [92]. For training SVM and KNN we used standard parameter settings available in *sci-kit learn* [93].

4.2 Single-task learning

We used several pre-trained models for single-task learning and fine-tuned the network with the task-specific classification layer on top of the network. This approach has been popular and has been performing well for various downstream visual recognition tasks [94–97]. The network architectures that we used in this study include ResNet18, ResNet50, ResNet101 [9], VGG16 [98], DenseNet [99], SqueezeNet [100], MobileNet [101], and EfficientNet [91]. We have chosen such diverse architectures to understand their relative performance and inference time. For fine-tuning, we use the weights of the networks pre-trained using ImageNet [102] to initialize our model. Our classification settings comprised binary (i.e., informativeness task) and multiclass settings (i.e., remaining three tasks). We train the models using the Adam optimizer [103] with an initial learning rate of 10^{-3}, which is decreased by a factor of 10 when accuracy on the dev set stops improving for 10 epochs. The models were trained for 150 epochs. We use the model with the best accuracy on the validation set to evaluate its performance on the test set.

4.3 Multi-task learning

In the MEDIC dataset, the tasks share similar properties; hence, we designed a simpler approach. We use the hard parameter sharing approach to reduce the computational complexity. All tasks share the same feature layers in the network, which is followed by task-specific classification layers. For optimizing the loss, we provide equal weight to each task. Assuming that the task-specific weight is w_i and task-specific loss function is $\mathcal{L}_i$, the optimization objective of the MTL is defined as $\mathcal{L}_{MTL} = \sum_i w_i.\mathcal{L}_i$. During optimization (i.e., using stochastic gradient descent to minimize the objective), the network weights in the shared layers W_{sh} are updated using the following equation:

$$\mathcal{W}_{sh} = \sum_i W_{sh} - \lambda \sum_i w_i \frac{\partial \mathcal{L}_i}{\partial W_{sh}} \tag{1}$$

We set $w_i = 1$ in our experiments for all task-specific weights, i.e., equal weight for all tasks. We use softmax activation to get probability distribution over individual tasks and use cross-entropy as a loss function. We initialized the weights using pre-trained models mentioned above, which are trained using ImageNet. Our implementation of multi-task learning supports all the network architectures mentioned in Sect. 4.2. Therefore, we have run experiments using the same pre-trained models and same hyper-parameter settings for the MTL experiments. We used the NVIDIA Tesla V100-SXM2-16 GB GPU machines consisting of 12 cores and 40GB CPU memory for all experiments.

4.4 Multilabel classification

In Table 5, we report the distribution of multilabel data split. It shows that a major part of the dataset is labeled with a single label for both tasks. For the multilabel classification, we run experiments in a single task learning

Table 7 Baseline classification results

Model	Acc	P	R	F1	Acc	P	R	F1
	Disaster Types				Informative			
Majority	56.6	32.1	56.6	41.0	45.9	21.1	45.9	28.9
Eff. Net Feat. + KNN	71.1	72.2	71.1	70.1	80.4	80.3	80.4	80.3
Eff. Net Feat. + SVM	75.7	74.1	75.7	**73.2**	83.0	83.0	83.0	**83.0**
	Humanitarian				Damage Severity			
Majority	58.3	34.0	58.3	42.9	65.3	42.7	65.3	51.7
Eff. Net Feat. + KNN	75.3	74.8	75.3	74.6	76.5	73.9	76.5	74.8
Eff. Net Feat. + SVM	77.9	76.1	77.9	**76.1**	78.3	75.1	78.3	**75.1**

Eff. Net Feat.: Feature extracted from the penultimate layer of a pre-trained efficient net model

Best F1 results are highlighted in bold

setup using the models mentioned above. We used the same training environment as other settings discussed in previous sections. However, we used sigmoid activation for multilabel instead of softmax, which is commonly used for multilabel setup.

4.5 Results

4.5.1 Baseline

In Table 7, we provide baseline results. From the majority baseline results it is clear that imbalance distribution does not play any role. Among SVM and KNN, the former is performing better in all tasks with 0.2 to 3.3% improvement.

4.5.2 Single- versus multi-task results

In Table 8, we report the results for both single- and multi-task settings using the mentioned models. Across different models, overall, EfficientNet (b1) performs better than other models. Comparing only EfficientNet (b1) results for all tasks, the multi-task setting shows better than single-task settings; although, the difference is minor and might

Table 8 Classification results using single and multi-task settings along with different pre-trained models

Model	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
	Single-task				Multi-task				Single-task				Multi-task			
	Disaster Types								Humanitarian							
ResNet18	79.8	78.3	79.8	78.1	79.8	79.1	79.8	77.9	82.6	81.6	82.6	81.9	83.2	82.0	83.2	82.2
ResNet50	80.6	79.7	80.6	79.0	80.9	80.0	80.9	79.4	83.4	83.1	83.4	83.0	84.2	83.5	84.2	83.7
ResNet101	81.3	80.4	81.3	79.6	81.1	81.0	81.1	78.9	83.9	83.1	83.9	83.4	84.6	83.7	84.6	83.9
VGG16	80.0	78.5	80.0	78.1	80.7	80.8	80.7	78.7	83.6	82.7	83.6	83.0	84.1	83.1	84.1	83.4
DenseNet (121)	81.1	80.2	81.1	79.5	80.7	80.2	80.7	78.8	83.4	82.5	83.4	82.7	83.9	83.0	83.9	83.2
SqueezeNet	76.5	75.0	76.5	73.6	77.1	75.5	77.1	74.7	79.8	78.0	79.8	78.4	81.0	79.5	81.0	79.9
MobileNet (v2)	80.1	79.0	80.1	78.0	79.9	79.2	79.9	78.0	82.7	81.7	82.7	82.0	83.5	82.5	83.5	82.7
EfficientNet (b1)	82.1	81.6	82.1	**80.7**	81.4	81.1	81.4	**79.8**	84.3	83.9	84.3	**84.0**	84.6	84.2	84.6	**84.3**
EfficientNet (b7)	81.0	79.9	81.0	79.1	80.5	79.5	80.5	78.7	83.2	82.4	83.2	82.7	83.2	82.4	83.2	82.7
	Informative								Damage Severity							
ResNet18	85.9	86.2	85.9	85.9	86.8	86.9	86.8	86.8	81.4	78.4	81.4	79.1	81.7	78.9	81.7	79.3
ResNet50	87.4	87.4	87.4	87.4	87.8	88.0	87.8	87.8	82.1	79.2	82.1	79.9	82.8	80.3	82.8	80.7
ResNet101	87.4	87.6	87.4	87.4	88.3	88.3	88.3	88.3	82.3	79.9	82.3	**80.6**	82.9	79.9	82.9	80.2
VGG16	86.7	87.1	86.7	86.8	87.6	87.7	87.6	87.6	82.3	79.6	82.3	79.7	82.7	80.1	82.7	80.5
DenseNet (121)	87.1	87.2	87.1	87.1	87.5	87.6	87.5	87.5	82.4	80.0	82.4	80.4	82.5	79.6	82.5	80.3
SqueezeNet	83.9	84.2	83.9	83.9	85.0	85.1	85.0	85.0	79.7	76.5	79.7	76.5	80.5	76.7	80.5	77.5
MobileNet (v2)	86.2	86.4	86.2	86.3	86.7	87.0	86.7	86.8	81.7	78.4	81.7	78.9	82.1	79.3	82.1	79.7
EfficientNet (b1)	87.7	87.7	87.7	**87.7**	88.6	88.7	88.6	**88.6**	82.8	80.3	82.8	80.4	82.9	80.7	82.9	**80.8**
EfficientNet (b7)	87.2	87.2	87.2	87.2	87.5	87.6	87.5	87.5	81.9	79.2	81.9	80.0	82.0	79.5	82.0	80.3

Best F1 result for each task setting is highlighted in bold

not be significant. However, since we share the feature layers across the four tasks, model space requirement and inference time are reduced by a factor of four. The improved inference time is crucial for real-time disaster response systems as it reduces the operational cost that running individual models would incur.

4.5.3 Multi-task results using different random seeds

In our experiment, only the weights of the last layer were initialized randomly, hence, this can result in a minor variation in the performance. We have run experiments using different random seeds with the MTL setting. In Table 9, we report results on selected models for all tasks. We observe that variation is very minor and among different models, DenseNet (121) shows relatively lower variation across tasks.

4.5.4 Ablation experiments in multi-task setup

To understand the task correlation and how they affect performance, we also run experiments with different subsets of the tasks (see Table 10). We obtain similar results with other task combinations. In Table 10, we show results obtained using combination of different subset of tasks. We observe that the results remain consistent with other combinations of tasks as well. It will be an important future research avenue to explore different weighting schemes for the tasks. Regardless, our reported results can serve as a baseline for single and multi-task disaster image classification.

4.5.5 Multilabel classification results

In Table 11, we report multilabel classification results for disaster types and humanitarian tasks. Overall, across

Table 9 Experiment using different random seeds in the MTL setup

Model	DT	Info	Hum	DS
ResNet101	79.4 ± 0.1	86.2 ± 0.1	80.7 ± 0.1	80.5 ± 0.1
VGG16	79.3 ± 0.4	86.2 ± 0.1	80.6 ± 0.1	80.3 ± 0.2
DenseNet (121)	79.2 ± 0.1	**86.2 ± 0.04**	**80.7 ± 0.03**	80.4 ± 0.1
EfficientNet (b1)	**79.5 ± 0.4**	88.5 ± 0.3	84.2 ± 0.1	**80.6 ± 0.3**

Table 10 Results (F1) with different combination of tasks using Efficient- Net (b1)

Model (task setup)	**DT**	Info	Hum	DS	Model (task setup)	DT	Info	Hum	DS
DT-Info-Hum-DS	79.8	88.6	84.3	80.8	DT-DS	80.7			**81.3**
DT-Info-Hum	80.3	**88.9**	**84.5**		Info-Hum-DS		88.3	84.0	80.8
DT-Info-DS	80.2	88.6		**81.0**	Info-Hum		**88.5**	83.9	
DT-Info	80.1	88.7			Info-DS		88.2		80.5
DT-Hum	**80.5**		84.4		Hum-DS			**84.1**	80.8

DT: Disaster type, Info: Informativeness, Hum: Humanitarian, DS: Damage severity

Table 11 Classification results using single-task multilabel settings with different pre-trained models

Model	Acc	P	R	F1	H	Acc	P	R	F1	H
	Disaster Type					Humanitarian				
ResNet18	73.9	86.2	73.2	79.2	6.2	76.5	85.4	78.4	81.8	9.6
ResNet50	76.3	86.1	75.6	80.5	5.9	78.6	86.4	80.5	83.3	8.8
ResNet101	75.8	86.2	75.9	80.7	5.9	79.0	86.6	80.4	**83.4**	8.8
VGG16	76.2	86.8	75.6	**80.8**	5.8	78.9	86.3	80.5	83.3	8.8
DenseNet (121)	75.8	87.6	74.3	80.4	5.9	78.0	86.5	78.9	82.5	9.1
SqueezeNet	36.3	41.8	64.3	50.7	20.3	31.9	55.3	78.9	65.0	23.2
MobileNet (v2)	73.5	86.4	71.8	78.4	6.4	76.8	86.0	78.0	81.8	9.5
EfficientNet (b1)	73.4	86.1	71.5	78.1	6.5	77.9	86.1	79.9	82.9	9.0
EfficientNet (b7)	76.0	86.0	74.7	80.0	6.1	78.2	85.4	80.3	82.8	9.1

H: Humming Loss lower is better. *Micro* average precision, recall, and F1

Table 12 Class-wise results for both single and multi-task settings using EfficientNet (b1) model

Label	P Single-task	R	F1	P Multi-task	R	F1
Disaster Type						
Earthquake	73.8	82.6	**77.9**	70.5	83.5	76.4
Fire	78.2	85.2	**81.6**	74.1	85.4	79.3
Flood	78.1	80.7	79.4	78.5	80.8	**79.6**
Hurricane	65.6	67.5	**66.6**	64.4	63.0	63.7
Landslide	62.2	78.5	**69.4**	60.4	75.5	67.1
Not disaster	88.9	92.9	**90.9**	88.9	92.7	90.8
Other disaster	70.5	18.8	**29.7**	72.6	15.6	25.7
Informativeness						
Informative	86.5	86.8	86.7	85.8	90.0	**87.9**
Not-informative	88.8	88.5	88.6	91.2	87.3	**89.2**
Humanitarian						
Affected, injured, or dead people	54.8	42.6	**47.9**	51.5	43.8	47.3
Infrastructure and utility damage	81.5	85.1	83.2	80.6	87.8	**84.1**
Not humanitarian	89.8	89.9	89.9	91.1	89.2	**90.1**
Rescue volunteering or donation effort	48.7	42.2	**45.2**	49.4	36.2	41.8
Damage Severity						
Litle or none	89.7	93.2	91.4	91.0	92.4	**91.7**
Mild	42.3	9.8	15.9	40.7	11.7	**18.2**
Severe	70.2	84.3	**76.6**	69.2	85.6	76.5

different models, SqueezeNet is the worst performing model, which we also observed for single and multi-task multiclass classification results. The multilabel results, as in Table 10, are not equally comparable with multiclass results, as reported in Table 8. This results will serve as baselines in future studies.

4.5.6 Error analysis

Given that class distribution can play a significant role in classifier performance, we explored whether low prevalent classes have any significant impact. In Table 12, we report task-wise classification results for both single and multi-task settings in which the model is trained using EfficientNet model. It appears that low prevalent classes have lower performance. However, this is not always the case. For example, the distribution of *Fire* class label is 3.8% in the dataset but the performance is third-best among class labels. Where the distribution of *Other disaster* is 5.1%, however, the F1 is 27.0, which is the lowest performance. With our analysis, we found that this *Other disaster* confused with *Not disaster* (see Table 14 in Appendix B).

In Tables 14, 15, 16 and 17 (in Appendix B) we report classification confusion matrices using EfficientNet (b1) model for disaster types, informative, humanitarian and damage severity, respectively. From the tables, we observe that there is comparable performances between different task settings. In some cases class label performance increases in multi-task setting and in some cases it decreases. For example, true positives increase for informative and decreases for not-informative in multi-task setting. The results in these tables also confirm the results in Table 8.

4.5.7 Computational time analysis

We have done extensive analysis to understand whether multi-task learning setup reduces computational time. In Table 13, we provide such findings for all the models we used in our experiments. From the results, it is clear that multi-task learning setup can significantly reduce the computation time both in terms of training and inference.

5 Discussion and future work

The MEDIC dataset provides images from diverse events consisting of different time frames. The crowdsourced annotation provides a reasonable annotator agreement even though the task is subjective. Our experiments show that multi-task learning with neural networks reduces

Table 13 Training and inference time in single- vs. multi-task settings with a batch size of 32. Time is in *day, hour:minute:second format*

Model	Single-task					Multi-task
	DT	Info	Hum	DS	Sum	All tasks
	Training time on the train set with 49,353 images					
ResNet18	21:38:48	17:15:09	16:53:02	17:41:23	3 days, 1:28:22	1 day, 3:41:02
ResNet50	21:14:49	17:16:03	21:41:07	17:24:00	3 days, 5:35:59	18:19:56
ResNet101	27:35:29	18:37:23	17:41:23	19:49:31	3 days, 11:43:46	1 day, 0:27:28
VGG16	19:53:52	23:43:49	23:15:04	23:37:10	3 days, 18:29:55	22:41:41
DenseNet (121)	20:23:39	17:08:27	17:23:06	18:21:06	3 days, 1:16:18	18:20:41
SqueezeNet	24:12:26	17:18:55	20:26:42	16:47:46	3 days, 6:45:49	18:12:44
MobileNet (v2)	17:44:03	21:39:41	17:55:16	21:06:44	3 days, 6:25:44	15:53:10
EfficientNet (b1)	21:59:19	17:37:01	17:28:30	17:08:27	3 days, 2:13:17	20:38:06
EfficientNet (b7)		26:39:17	26:40:33	26:55:17	3 days, 8:15:07	1 day, 16:13:38
	Inference time on the test set with 15,688 images					
ResNet18	0:02:26	0:01:56	0:05:11	0:01:53	0:11:26	0:05:10
ResNet50	0:02:25	0:01:55	0:02:24	0:01:54	0:08:38	0:02:13
ResNet101	0:05:20	0:07:48	0:02:05	0:02:08	0:17:21	0:01:58
VGG16	0:05:21	0:01:57	0:05:10	0:01:56	0:14:24	0:02:15
DenseNet (121)	0:02:08	0:01:55	0:01:57	0:05:22	0:11:22	0:02:08
SqueezeNet	0:10:59	0:01:54	0:02:22	0:05:15	0:20:30	0:04:44
MobileNet (v2)	0:01:57	0:02:26	0:01:56	0:02:26	0:08:45	0:01:57
EfficientNet (b1)	0:05:17	0:01:56	0:02:07	0:01:54	0:11:14	0:02:32
EfficientNet (b7)		0:02:12	0:02:11	0:02:10	0:06:33	0:02:13

computational complexity significantly while having comparative performance.

In Fig. 2, we show the loss and accuracy plots for single and multi-task settings for EfficientNet (b1) model. We limit the plots to 40 epochs as all of the models converged by then. We notice similar convergence rates for both single and multi-task learning setups. We observe that the multi-task objective function acts as a regularizer as the training loss is consistently higher and training accuracy is lower than the single-task setting while having similar or better performance on the validation set. This suggests that the multi-task setup may benefit from models having a larger capacity.

Class distribution is an important issue that affect classifier performance. We investigated class-wise performances and confusion matrix. Our observation suggests that imbalanced class distribution is not only factor for lower classification performance in certain classes. It also depends on distinguishing properties of the class label. For example, the distribution of *Fire* class label is 3.8% in the dataset but the performance is third-best among class labels. Where the distribution of *Other disaster* is 5.1%, however, the F1 is 27.0, which is the lowest performance.

Future work

Our future work includes exploring other multi-task learning methods, and investigating tasks groups and relationships. For instance, further investigation is needed to explain why training the model with disaster types, informativeness and humanitarian tasks reduces performance as presented in Table 10. Other research avenues include multimodality (e.g., integrating text), and investigating class imbalance issues.

6 Conclusions

We presented a large-scale, manually annotated multi-task learning dataset, comprising 71,198 images labeled for four tasks, which were specifically designed for multi-task learning research and disaster response image classification. The dataset will not only be useful to develop robust models for disaster response tasks but will also enable evaluation of general multi-task models. We provide classification results using nine different pre-trained models, which can serve as a benchmark for future work. We report that the multi-task model reduces the inference time significantly, hence, such a model can be very useful for real-time classification tasks, especially for analyzing social media image streams.

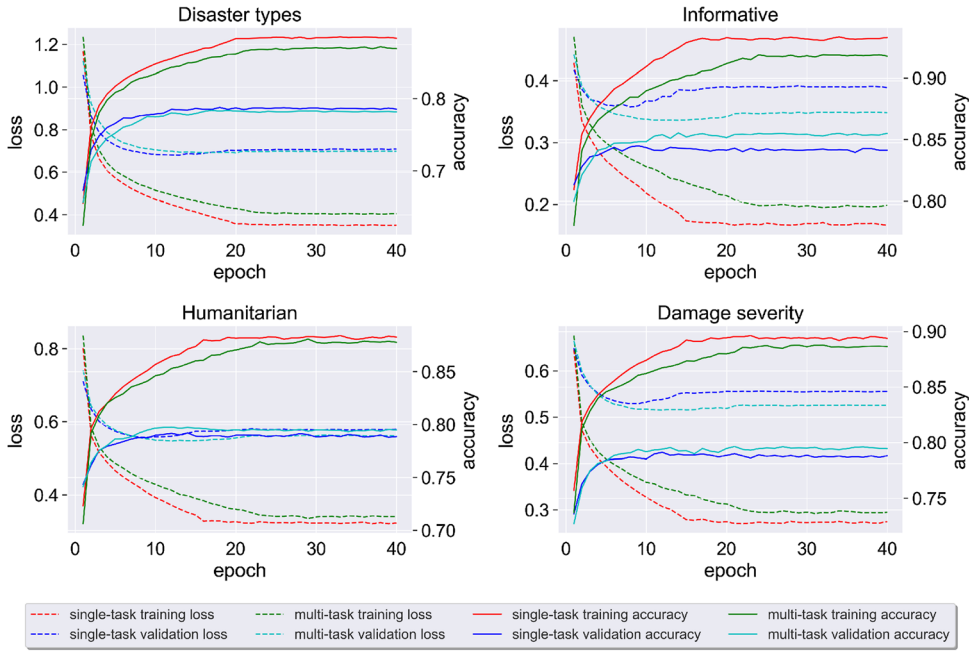


Fig. 2 Training and validation loss and accuracy for EfficientNet (b1) model for single and multi-task settings

Appendix

Appendix A Data collection

A.1 Data curation and annotation

We extended the Crisis Benchmark dataset to develop MEDIC, a multi-task learning dataset for disaster response. For the annotation, we provided detailed instructions to the annotators, which they followed during the annotation tasks. Our annotation consists of four tasks in different batches, and we provided task-specific instructions along with them.

A.2 Annotation instructions

The annotation task involves identifying images that are useful for humanitarian aid/response. During different disaster events (i.e., natural and human-induced or hybrid), *humanitarian aid* involves assisting people who need help. The primary purpose of humanitarian aid is to save lives, reduce suffering, and rebuild affected communities. Among the people in need belong homeless, refugees, and victims of natural disasters, wars, and conflicts who need necessities like food, water, shelter, medical assistance, and damage-free critical infrastructure and utilities such as roads, bridges, power lines, and communication poles.

For disaster types and humanitarian tasks, it is possible that some images can be annotated with multiple labels. In such cases, the instruction is to choose a label that is critical (i.e., higher priority) for humanitarian organizations and more prominent in the image.

A.2.1 Disaster types

The purpose of identifying disaster type is to understand the type of disaster events shared in an image. The annotation task involves looking into the image can carefully select one of the following disaster types based on their specific definition. There might be the case that an image shows an effect of a hurricane (destroyed house) and also flood, in such cases the task is to carefully check what is more visible and select label accordingly. Example of images demonstrating different disaster types is shown in Fig. 3.

- *Earthquake:* this type of images shows damaged or destroyed buildings, fractured houses, ground ruptures such as railway lines, roads, airport runways, highways, bridges, and tunnels.
- *Fire:* image shows man-made fires or wildfires (forests, grasslands, brush, and deserts), destroyed forests, houses, or infrastructures.
- *Flood:* image shows flooded areas, houses, roads, and other infrastructures.
- *Hurricane:* image shows high winds, a storm surge, heavy rains, collapsed electricity polls, grids, and trees.

Fig. 3 Examples of images *disaster types*

Fig. 4 Example images for *informativeness*

- *Landslide:* image shows landslide, mudslide, landslip, rockfall, rockslide, earth slip, and land collapse
- *Other disasters:* image shows any other disaster types such as plane crash, bus, car, or train accident, explosion, war, and conflicts.
- *Not disaster:* image shows cartoon, advertisement, or anything that cannot be easily linked to any disaster type.

A.2.2 Informativeness

The purpose of this task is to determine whether image is useful for *humanitarian aid* purposes as defined below. If the given image is useful for *humanitarian aid*, the annotation task is to select the label "Informative", otherwise select the label "Not informative" image. Example of images demonstrating informative versus not-informative is shown in Figs. 3, 4.

- *Informative:* if an image is useful for humanitarian aid and shows one or more of the following: cautions, advice, and warnings, injured, dead, or affected people, rescue, volunteering, or donation request or effort, damaged houses, damaged roads, damaged buildings; flooded houses, flooded streets; blocked roads, blocked bridges, blocked pathways; any built structure affected by earthquake, fire, heavy rain, strong winds, gust, etc., disaster area maps.
- *Not informative:* if the image is not useful for humanitarian aid and shows advertising, banners, logos, cartoons, and blurred.

A.2.3 Humanitarian categories

Based on the *humanitarian aid* definition above, we define each *humanitarian* information category below. Example images are shown in Fig. 5.

- *Affected, injured or dead people:* image shows injured, dead, or affected people such as people in shelter facilities, sitting or lying outside, etc.
- *Infrastructure and utility damage:* image shows any built structure affected or damaged by the disaster. This includes damaged houses, roads, buildings; flooded houses, streets, highways; blocked roads, bridges,

Fig. 5 Example images for *humanitarian* categories

Fig. 6 Example images for *damage severity*

pathways; collapsed bridges, power lines, communication poles, etc.

- *Not humanitarian:* image is not relevant or useful for humanitarian aid and response such as non-disaster scenes, cartoons, advertisement banners, celebrities, etc.
- *Rescue, volunteering, or donation effort:* image shows any type of rescue, volunteering, or response effort such as people being transported to safe places, people being evacuated from the hazardous area, people receiving medical aid or food, donation of money, blood, or services, etc.

A.2.4 Damage Severity

The purpose of this task is to identify the severity of damage reported in an image (Fig. 6). It can be physical destruction to a build-structure. Our goal is to detect physical damages like broken bridges, collapsed or shattered buildings, destroyed or creaked roads. We define each damage severity category below.

1. *Severe:* Substantial destruction of an infrastructure belongs to the severe damage category. For example, a non-livable or non-usable building, a non-crossable bridge, or a non-drivable road, destroyed, burned crops, forests are all examples of severely damaged infrastructures. For example, if one or more building in the image show substantial loss of amenity or images shows a building that is not safe to use then such image should be labeled as severe damage.
2. *Mild:* Partially destroyed buildings, bridges, houses, roads belong to mild damage category. For example, if image shows a building with damage up to 50%, partial loss of amenity/roof or part of the building can has to be closed down then it should label as mild damage.
3. *Little or none:* Images that show damage-free infrastructure (except for wear and tear due to age or disrepair) belong to the little-or-no-damage category.

A.3 Annotation interface

An example of annotation interface is show in in Fig. 7. Image on the left shows annotation task is launched to annotate image for disaster type and humanitarian tasks and image on the right shows annotation task is launched for three tasks.

A.4 Manual annotation

In our annotation tasks through the Appen platform, more than 3000 annotators participated from more than 50 countries. For the annotation task, we estimated hourly wages and it was 6 to 8 USD per hour on average, which varied depending on the two to three labels annotation per image. We think such pay is reasonable as annotators are from various part of the world where wages varies depending on the location. In total we paid 5,159 USD for the annotation, including Appen charges.

Fig. 7 Example of annotation interfaces on Appen crowdsoursing platform. DT: disaster type, Hum: humanitarian, DS: damage severity

A.5 Data analysis

In Fig. 8, we report class-wise relationship between tasks. It appears that there is an association between labels for different tasks. For example, for disaster types and informativeness tasks, as shown in the Fig. 8a, *not disaster* and *not informative* are highly related. A major part of *not disaster* images are labeled as *little or none* damages as shown in 8b. Our observations for other task combinations are quite similar for different label pairs.

Appendix B Error analysis

In Tables 14, 15, 16 and 17, we report confusion matrices for different tasks with a comparison to single vs. multi-task settings.

Model Parameters

In Table 18, we report model parameters and their memory consumption.

Appendix C The MEDIC dataset

The dataset can be downloaded from https://crisisnlp.qcri.org/medic/index.html.

C.1 Data format

The dataset format can be found in https://crisisnlp.qcri.org/medic/index.html.

C.2 Terms of use, privacy and license

The MEDIC dataset is published under CC BY-NC-SA 4.0 license, which means everyone can use this dataset for non-commercial research purpose: https://creativecommons.org/licenses/by-nc/4.0/.

C.3 Data maintenance

We provide data download link through https://crisisnlp.qcri.org/medic/index.html. We also host the dataset on Zenodo[8] for wider access. We will maintain the data for a long period of time and make sure dataset is accessible.

C.4 Benchmark code

The benchmark code is available at: https://github.com/firojalam/medic/.

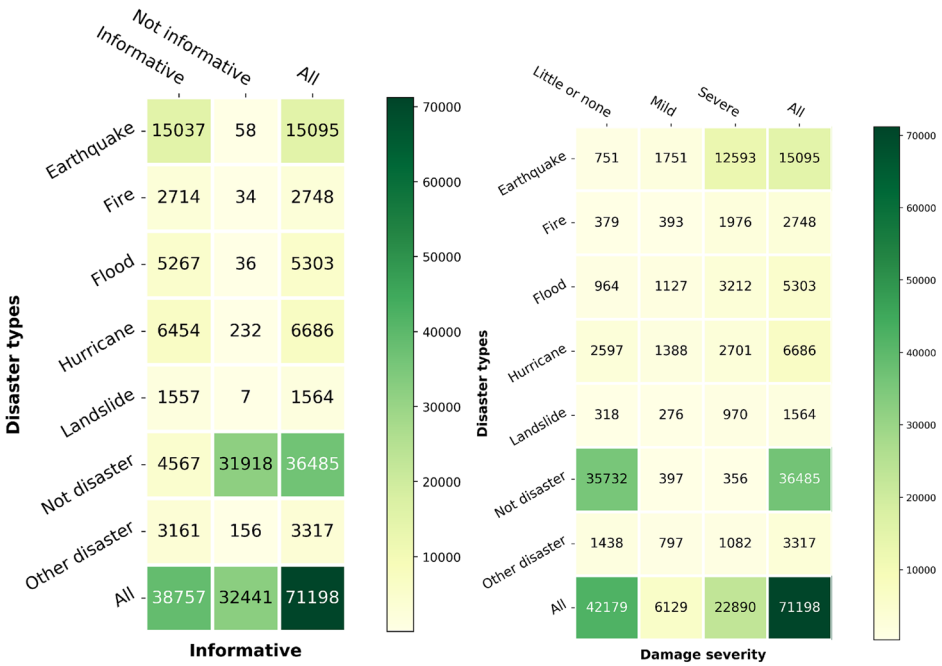
C.5 Ethics statement

C.5.1 Dataset collection

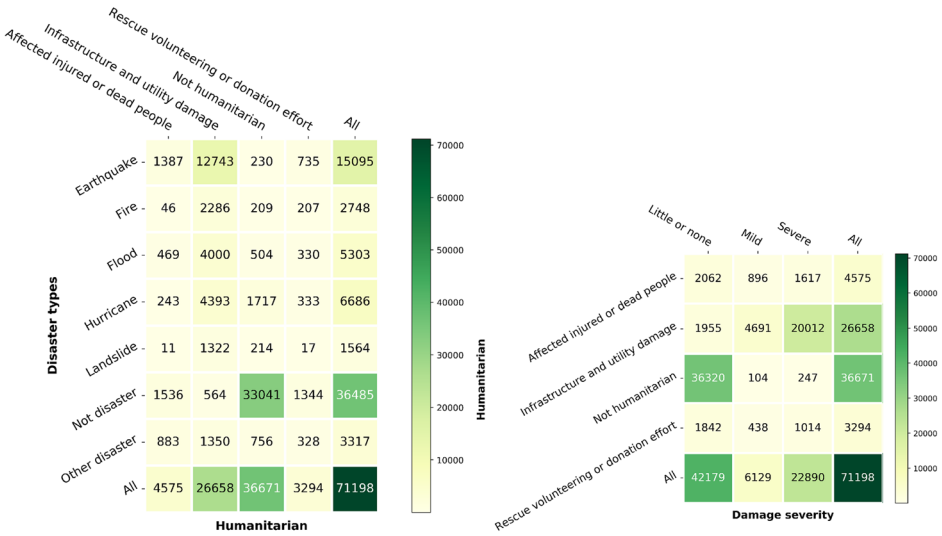
The dataset contains images from multiple sources such as Twitter, Google, Bing, Flickr, and Instagram. Twitter developer terms and conditions suggests that one can release 50K tweet objects[9] and here we only provide images not whole JSON objects. The total number of images from Twitter is less than 50,000. Hence, by

[8] https://doi.org/10.5281/zenodo.6625120.

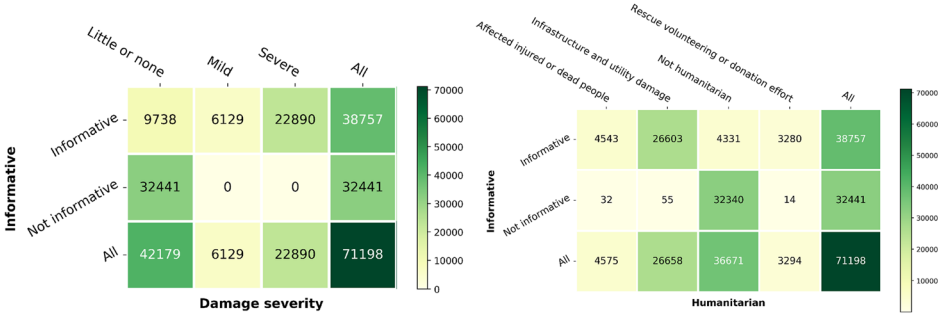
[9] http://developer.twitter.com/en/developer-terms/agreement-and-policy.



(a) DT and info. **(b)** DT and DS.



(c) DS and hum. **(d)** Hum and DS.



(e) Info and DS. **(f)** Info and hum.

Fig. 8 Contingency heatmaps for different pairs of tasks

Table 14 Confusion matrix for *disaster types* task using single vs. multi-task learning with efficient-net (b1) model

Label	Earthquake	Fire	Flood	Hurricane	Landslide	Not disaster	Other disaster	Total
Single-task								
Earthquake	1482	22	14	66	38	156	17	1795
Fire	17	588	3	9	4	66	3	690
Flood	19	5	1061	64	20	145	1	1315
Hurricane	104	10	92	1025	29	234	24	1518
Landslide	27	3	7	13	260	21	0	331
Not disaster	122	53	142	241	28	8253	46	8885
Other disaster	237	71	39	144	39	407	217	1154
Total	2008	752	1358	1562	418	9282	308	15688
Multi-task								
Earthquake	1498	22	12	69	34	150	10	1795
Fire	21	589	3	11	3	60	3	690
Flood	24	6	1062	55	20	147	1	1315
Hurricane	151	16	112	956	41	232	10	1518
Landslide	30	4	5	16	250	26	0	331
Not disaster	130	76	121	243	34	8237	44	8885
Other disaster	272	82	38	135	32	415	180	1154
Total	**2126**	**795**	**1353**	**1485**	**414**	**9267**	**248**	**15688**

Table 15 Confusion matrix for *informative* task using single vs. multi-task learning with efficient-net (b1) model

	Single-task			Multi-task		
Label	Informative	Not Informative	Total	Informative	Not Informative	Total
Informative	6256	950	7206	6489	717	7206
Not Informative	977	7505	8482	1076	7406	8482
Total	**7233**	**8455**	**15688**	**7565**	**8123**	**15688**

Table 16 Confusion matrix for *humanitarian* task using single vs. multi-task learning with efficient-net (b1) model

Label	Affected	Infra. damage	Not hum	Rescue	Total
Single-task					
Affected, injured, or dead people	272	181	149	37	639
Infrastructure and utility damage	68	4445	630	81	5224
Not humanitarian	93	649	8219	184	9145
Rescue volunteering or donation effort	63	180	150	287	680
Total	496	5455	9148	589	15688
Multi-task					
Affected, injured, or dead people	280	191	137	31	639
Infrastructure and utility damage	67	4588	522	47	5224
Not humanitarian	118	698	8155	174	9145
Rescue volunteering or donation effort	79	214	141	246	680
Total	**544**	**5691**	**8955**	**498**	**15688**

Table 17 Confusion matrix for *damage severity* task using single vs. multi-task learning with efficient-net (b1) model

	Single-task				Mult-task			
Label	Little or none	Mild	Severe	Total	Little or none	Mild	Severe	Total
Litle or none	9550	120	582	10252	9476	152	624	10252
Mild	563	149	815	1527	481	179	867	1527
Severe	530	83	3296	3909	453	109	3347	3909
Total	**10643**	**352**	**4693**	**15688**	**10410**	**440**	**4838**	**15688**

Table 18 Model parameters: number of layer, parameter, and size of the memory consumption

Arch	# Layer	# Param (M)	Memory (MB)
ResNet18	18	11.18	74.61
ResNet50	50	23.51	233.54
ResNet101	101	42.50	377.58
AlexNet	8	57.01	222.24
VGG16	16	134.28	673.87
DenseNet (121)	121	6.96	174.2
SqueezeNet	18	0.74	47.99
InceptionNet (v3)	42	24.35	206.01
MobileNet (v2)	20	2.23	8.49
EfficientNet (b1)	25	7.79	177.82

releasing the data by maintaining such terms and conditions. From Google, Bing, Yahoo and Instagram images are publicly available. In addition, we also maintain licenses and cite prior work based upon we built our work.

C.5.2 Potential negative societal impacts

The dataset consists of images collected from social media and different search engines. We have given our best efforts to eliminate any adult content during data preparation and annotation. Hence, we believe that the presence of such content in the dataset might be very unlikely. Our annotation does not contain any identifiable information such as age, gender, or race. However, the images in the dataset have many faces and one might apply facial recognition to identify someone. Intervention with human moderation would be required in order to ensure this does not lead to any misuse. We also would like to highlight that the models' prediction should be used carefully as the purpose of the models' prediction is to facilitate its user, not to make any direct decision. Model designers also need to be careful for any adversarial attack that can lead to creation and spread of any mis/disinformation.

C.5.3 Biases

The datasets are not representative of a geolocation, user gender, age, race, so should not be used in analyses requiring a representative sample. Instead, the datasets are more suitable to be combined with existing datasets and used for training supervised machine learning models.

We also would like to highlight that some of the annotations are subjective, and we have clearly indicated in the text which of these are. Thus, it is inevitable that there would be biases in our dataset. Note that, we have very clear annotation instructions with examples in order to reduce such biases.

C.5.4 Intended use

The dataset can enable an analysis of image content for disaster response, which could be of interest to crisis responders humanitarian response organizations, and policymakers. There are only very few datasets available for multi-task learning research. This dataset can significantly help towards this direction. Having a single model for multiple tasks can also foster Green AI.

Funding Open Access funding provided by the Qatar National Library.

Data availability The dataset proposed in this research is available to download from the following links: https://crisisnlp.qcri.org/medic/index.html and https://doi.org/10.5281/zenodo.6625120. More details about the dataset is available in Appendix C.

Declarations

Conflict od interest The authors have no competing interests.

References

1. Mouzannar H, Rizk Y, Awad M (2018) Damage identification in social media posts using multimodal deep learning. In: Proceedings of the international conference on information systems for crisis response and management. ISCRAM '18. ISCRAM Association, pp 529–543
2. Nguyen DT, Ofli F, Imran M, 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. ASONAM '17. IEEE, pp 1–8
3. Alam F, Ofli F, Imran M (2018) CrisisMMD: multimodal twitter datasets from natural disasters. In: Proceedings of the International AAAI conference on web and social media. ICWSM '18. AAAI, pp 465–473
4. Weber E, Marzo N, Papadopoulos DP, Biswas A, Lapedriza A, Ofli F, Imran M, Torralba A (2020) Detecting natural disasters, damage, and incidents in the wild. In: Proceedings of the European conference on computer vision. ECCV '20. Springer, pp 331–350
5. Alam F, Ofli F, Imran M, Alam T, Qazi U (2020) Deep learning benchmarks and datasets for social media image classification for disaster response. In: Proceedings of the IEEE/ACM international conference on advances in social networks analysis and mining. ASONAM '20. IEEE, pp 151–158. https://doi.org/10.1109/ASONAM49781.2020.9381294
6. Imran M, Castillo C, Diaz F, Vieweg S (2015) Processing social media messages in mass emergency: a survey. ACM Comput Surv 47(4):67
7. Said N, Ahmad K, Riegler M, Pogorelov K, Hassan L, Ahmad N, Conci N (2019) Natural disasters detection in social media and satellite imagery: a survey. Multimedia Tools Appl 78(22):31267–31302
8. Imran M, Ofli F, Caragea D, Torralba A (2020) Using ai and social media multimodal content for disaster response and management: opportunities, challenges, and future directions. Inf Process Manag 57(5):102261. https://doi.org/10.1016/j.ipm.2020.102261
9. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. CVPR '16. IEEE, pp 770–778
10. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. CVPR '15. IEEE, pp 3431–3440
11. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. CVPR '16. IEEE, pp 779–788
12. Alam F, Ofli F, Imran M (2018) Processing social media images by combining human and machine computing during crises. Int J Hum Comput Interact 34(4):311–327. https://doi.org/10.1080/10447318.2018.1427831
13. Yu F, Chen H, Wang X, Xian W, Chen Y, Liu F, Madhavan V, Darrell T (2020) BDD100k: a diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. CVPR '20. IEEE, pp 2636–2645
14. Caruana R (1997) Multitask learning. Mach Learn 28(1):41–75
15. Vandenhende S, Georgoulis S, Van Gansbeke W, Proesmans M, Dai D, Van Gool L (2021) Multi-task learning for dense prediction tasks: a survey. IEEE Trans Pattern Anal Mach Intell
16. Alam F, Imran M, Ofli F (2017) Image4Act: online social media image processing for disaster response. In: Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining. ASONAM '17. IEEE, pp 1–4
17. Schwartz R, Dodge J, Smith NA, Etzioni O (2020) Green AI. Commun ACM 63(12):54–63
18. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2015) ImageNet large scale visual recognition challenge. Int J Comput Vis 115(3):211–252. https://doi.org/10.1007/s11263-015-0816-y
19. Lin T, Maire M, Belongie SJ, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft COCO: common objects in context. In: Proceedings of the European conference on computer vision. ECCV '14, pp 740–755
20. Ruder S (2017) An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098
21. Zhang Y, Yang Q (2021) A survey on multi-task learning. IEEE Trans Knowl Data Eng
22. Crawshaw M (2020) Multi-task learning with deep neural networks: a survey. arXiv preprint arXiv:2009.09796
23. Worsham J, Kalita J (2020) Multi-task learning for natural language processing in the 2020s: where are we going? Pattern Recogn Lett 136:120–126
24. Strezoski G, van Noord N, Worring M (2019) Learning task relatedness in multi-task learning for images in context. In: Proceedings of the 2019 on International conference on multimedia retrieval. pp 78–86
25. Kokkinos I (2017) Ubernet: training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In: Proceedings of the IEEE conference on computer vision and pattern recognition. CVPR '17. IEEE, pp 6129–6138
26. Kendall A, Gal Y, Cipolla R (2018) Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE conference on computer vision and pattern recognition. CVPR '18. IEEE, pp 7482–749
27. Chen Z, Badrinarayanan V, Lee C-Y, Rabinovich A (2018) Gradnorm: gradient normalization for adaptive loss balancing in deep multitask networks. In: Proceedings of the international conference on machine learning. PMLR, pp 794–803
28. Misra I, Shrivastava A, Gupta A, Hebert M (2016) Cross-stitch networks for multi-task learning. In: Proceedings of the IEEE Conference on computer vision and pattern recognition. pp 3994–4003
29. Ruder S, Bingel J, Augenstein I, Søgaard A (2019) Latent multi-task architecture learning. In: Proceedings of the AAAI conference on artificial intelligence. AAAI '19, vol. 33. AAAI, pp 4822–4829
30. Gao Y, Ma J, Zhao M, Liu W, Yuille AL (2019) Nddr-CNN: layerwise feature fusing in multi-task cnns by neural discriminative dimensionality reduction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. CVPR '19. IEEE, pp 3205–3214
31. Yang Y, Hospedales T (2017) Deep multi-task representation learning: a tensor factorisation approach. In: Proceedings of the 5th International conference on learning representations

32. Kang Z, Grauman K, Sha F (2011) Learning with whom to share in multi-task feature learning. In: International conference on machine learning
33. Kumar A, Daumé III H (2012) Learning task grouping and overlap in multi-task learning. In: Proceedings of the 29th International coference on international conference on machine learning. pp 1723–1730
34. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324
35. Hull JJ (1994) A database for handwritten text recognition research. IEEE Trans Pattern Anal Mach Intell 16(5):550–554
36. Gong B, Shi Y, Sha F, Grauman K (2012) Geodesic flow kernel for unsupervised domain adaptation. In: Proceedings of the 2012 IEEE Conference on computer vision and pattern recognition. CVPR '12. IEEE, pp 2066–2073
37. Yang Y, Hospedales T (2016) Deep multi-task representation learning: a tensor factorisation approach. arXiv preprint arXiv:1605.06391
38. Liu Z, Luo P, Wang X, Tang X (2015) Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision. pp 3730–3738
39. Eidinger E, Enbar R, Hassner T (2014) Age and gender estimation of unfiltered faces. IEEE Trans Inf Forensics Secur 9(12):2170–2179
40. Strezoski G, Worring M (2018) Omniart: a large-scale artistic benchmark. ACM Trans Multimedia Comput Commun Appl (TOMM) 14(4):1–21
41. Silberman N, Hoiem D, Kohli P, Fergus R (2012) Indoor segmentation and support inference from rgbd images. In: Proceedings of the European conference on computer vision. Springer, pp 746–760
42. Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A (2010) The pascal visual object classes (voc) challenge. Int J Comput Vis 88(2):303–338
43. Chen X, Mottaghi R, Liu X, Fidler S, Urtasun R, Yuille A (2014) Detect what you can: detecting and representing objects using holistic models and body parts. In: Proceedings of the IEEE conference on computer vision and pattern recognition. CVPR '14. IEEE, pp 1971–1978
44. Zamir AR, Sax A, Shen W, Guibas LJ, Malik J, Savarese S (2018) Taskonomy: disentangling task transfer learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 3712–3722
45. Krizhevsky A, Hinton G, et al. (2009) Learning multiple layers of features from tiny images. Master's thesis, Department of Computer Science, University of Toronto
46. Wah C, Branson S, Welinder P, Perona P, Belongie S (2011) The caltech-ucsd birds-200-2011 dataset
47. Lake BM, Salakhutdinov R, Tenenbaum JB (2015) Human-level concept learning through probabilistic program induction. Science 350(6266):1332–1338
48. Wah C, Branson S, Welinder P, Perona P, Belongie S (2011) The caltech-ucsd birds-200-2011 dataset
49. Venkateswara H, Eusebio J, Chakraborty S, Panchanathan S (2017) Deep hashing network for unsupervised domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. CVPR '17. IEEE, pp 5018–5027
50. Lampert CH, Nickisch H, Harmeling S (2009) Learning to detect unseen object classes by between-class attribute transfer. In: 2009 IEEE conference on computer vision and pattern recognition. IEEE, pp 951–958
51. Gupta R, Goodman B, Patel N, Hosfelt R, Sajeev S, Heim E, Doshi J, Lucas K, Choset H, Gaston M (2019) Creating xbd: a dataset for assessing building damage from satellite imagery. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition workshops. CVPR '19. IEEE
52. Benjamin B, Patrick H, Zhengyu Z, de BJ, Damian B (2018) The multimedia satellite task at MediaEval 2018: emergency response for flooding events. In: Proceedings of the MediaEval
53. Nguyen DT, Alam F, Ofli F, Imran M (2017) Automatic image filtering on social networks using deep learning and perceptual hashing during crises. In: Proceedings of the information systems for crisis response and management. ISCRAM '17. ISCRAM Association
54. Bischke B, Helber P, Schulze C, Srinivasan V, Dengel A, Borth D (2017) The multimedia satellite task at MediaEval 2017. In: Proceedings of the MediaEval 2017: MediaEval Benchmark Workshop
55. Imran M, Qazi U, Ofli F, Peterson S, Alam F (2022) Ai for disaster rapid damage assessment from microblogs. In: Thirty-fourth annual conference on innovative applications of artificial intelligence (IAAI-22)
56. Alam F, Joty S, Imran M (2018) Domain adaptation with adversarial training and graph embeddings. In: Proceedings of the 56th annual meeting of the association for computational linguistics. ACL. Association for Computational Linguistics, Melbourne, Australia, pp 1077–1087 https://doi.org/10.18653/v1/P18-1099
57. Alam F, Muhammad I, Ferda O (2019) Crisisdps: crisis data processing services. In: Proceedings of the international conference on information systems for crisis response and management (ISCRAM)
58. Imran M, Castillo C, Lucas J, Meier P, Vieweg S (2014) AIDR: artificial intelligence for disaster response. In: Proceedings of the 23rd International conference on world wide web. pp 159–162
59. Villegas C, Martinez M, Krause M (2018) Lessons from harvey: drisis informatics for urban resilience. In: Rice University Kinder Institute for Urban Research
60. Olteanu A, Castillo C, Diaz F, Vieweg S (2014) Crisislex: a lexicon for collecting and filtering microblogged communications in crises. In: Proc. of ICWSM
61. Imran M, Mitra P, Castillo C (2016) Twitter as a lifeline: human-annotated twitter corpora for nlp of crisis-related messages. In: Proceedings of the international conference on language resources and evaluation. ELRA, Paris, France
62. Mccreadie R, Buntain C, Soboroff I (2019) Trec incident streams: finding actionable information on social media. In: Proceedings of the international conference on information systems for crisis response and management (ISCRAM)
63. Wiegmann M, Kersten J, Klan F, Potthast M, Stein B (2020) Analysis of detection models for disaster-related tweets. In: Proceedings of the 17th international conference on information systems for crisis response and management. ISCRAM '20. ISCRAM Association. https://doi.org/10.5281/zenodo.3713920
64. Alharbi A, Lee M (2019) Crisis detection from Arabic tweets. In: Proceedings of the 3rd workshop on Arabic corpus linguistics. pp 72–79
65. Alam F, Sajjad H, Imran M, Ofli F (2021) CrisisBench: benchmarking crisis-related social media datasets for humanitarian information processing. In: Proceedings of the International AAAI conference on web and social media. ICWSM '21, vol. 15. AAAI, pp 923–932
66. Alam F, Qazi U, Imran M, Ofli F (2021) HumAID: human-annotated disaster incidents data from twitter with deep learning benchmarks. In: Proceedings of the Fifteenth International AAAI conference on web and social media. ICWSM '21. AAAI, pp 933–942
67. Ofli F, Alam F, Imran M (2020) Analysis of social media data using multimodal deep learning for disaster response. In:

Proceedings of the 17th international conference on information systems for crisis response and management. ISCRAM '20. ISCRAM Association
68. Burel G, Saif H, Fernandez M, Alani H (2017) On semantics and deep learning for event detection in crisis situations. In: Workshop on Semantic Deep Learning (SemDeep), at ESWC 2017
69. Burel G, Alani H (2018) Crisis event extraction service (crees)-automatic detection and classification of crisis-related content on social media. In: Proceedings of the international conference on information systems for crisis response and management (ISCRAM)
70. Imran M, Alam F, Qazi U, Peterson S, Ofli F (2020) Rapid damage assessment using social media images by combining human and machine intelligence. In: Proceedings of the 17th international conference on information systems for crisis response and management. ISCRAM '20. ISCRAM Association, pp 761–773
71. Peters R, de Albuqerque JP (2015) Investigating images as indicators for relevant social media messages in disaster management. In: Proceedings of the international conference on information systems for crisis response and management. ISCRAM '15. ISCRAM Association
72. Daly S, Thom J (2016) Mining and classifying image posts on social media to analyse fires. In: Proceedings of the international conference on information systems for crisis response and management. ISCRAM '16. ISCRAM Association, pp 1–14
73. Nia KR, Mori G (2017) Building damage assessment using deep learning and ground-level image data. In: Proceeding of the 14th conference on computer and robot vision (CRV). IEEE, pp 95–102
74. Li X, Caragea D, Zhang H, Imran M (2018) Localizing and quantifying damage in social media images. In: Proceeding of the 2018 IEEE/ACM international conference on advances in social networks analysis and mining. ASONAM '18, pp. 194–201. IEEE
75. Li X, Caragea D, Caragea C, Imran M, Ofli F (2019) Identifying disaster damage images using a domain adaptation approach. In: Proceeding of the international conference on information systems for crisis response and management. ISCRAM '19. ISCRAM Association, pp 633–645
76. Pouyanfar S, Tao Y, Sadiq S, Tian H, Tu Y, Wang T, Chen S-C, Shyu M-L (2019) Unconstrained flood event detection using adversarial data augmentation. In: Proceeding of the IEEE international conference on image processing. ICIP '19. IEEE, pp 155–159
77. Ahmad S, Ahmad K, Ahmad N, Conci N (2017) Convolutional neural networks for disaster images retrieval. In: Proceedings of the MediaEval
78. Lagerstrom R, Arzhaeva Y, Szul P, Obst O, Power R, Robinson B, Bednarz T (2016) Image classification to support emergency situation awareness. Front Robot AI 3:54. https://doi.org/10.3389/frobt.2016.00054
79. Ning H, Li Z, Hodgson ME et al (2020) Prototyping a social media flooding photo screening system based on deep learning. ISPRS Int J Geo Inf 9(2):104
80. Hassan SZ, Ahmad K, Al-Fuqaha A, Conci N (2019) Sentiment analysis from images of natural disasters. In: Proceedings of the international conference on image analysis and processing. Springer, pp 104–113
81. Ahmad K, Riegler M, Pogorelov K, Conci N, Halvorsen P, De Natale F (2017) Jord: a system for collecting information and monitoring natural disasters by linking social media with satellite imagery. In: Proceedings of the 15th international workshop on content-based multimedia indexing. pp 1–6
82. Jony RI, Woodley A, Perrin D (2019) Flood detection in social media images using visual features and metadata. Proceedings of the 2019 Digital Image Computing: Techniques and Applications. IEEE, pp 1–8
83. Shaluf IM (2007) Disaster types. Disaster Prev Manag Int J
84. Chowdhury SA, Calvo M, Ghosh A, Stepanov EA, Bayer AO, Riccardi G, García F, Sanchis E (2015) Selection and aggregation techniques for crowdsourced semantic annotation task. In: Proceedings of the sixteenth annual conference of the international speech communication association. ISCA
85. Fleiss JL, Levin B, Paik MC (2013) Statistical methods for rates and proportions
86. Krippendorff K (1970) Estimating the reliability, systematic error and random error of interval data. Educ Psychol Meas 30(1):61–70
87. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 159–174
88. Alam F, Alam T, Imran M, Ofli F (2021) Robust training of social media image classification models for rapid disaster response. arXiv:2104.04184 [cs.CV]
89. Wu X-Z, Zhou Z-H (2017) A unified view of multi-label performance measures. In: International Conference on machine learning. PMLR, pp 3780–3788
90. Sorower MS (2010) A literature survey on algorithms for multi-label learning. Oregon State University, Corvallis 18:1–25
91. Tan M, Le QV (2019) Efficientnet: rethinking model scaling for convolutional neural networks. arXiv:1905.11946
92. Nakov P, Da San Martino G, Elsayed T, Barrón-Cedeño A, Míguez R, Shaar S, Alam F, Haouari F, Hasanain M, Mansour W, Hamdan B, Ali ZS, Babulkov N, Nikolov A, Shahi GK, Struß JM, Mandl T, Kutlu M, Kartal YS (2021) Overview of the CLEF-2021 CheckThat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In: Candan, K., Ionescu, B., Goeuriot, L., Larsen, B., Müller, H., Joly, A., Maistro, M., Piroi, F., Faggioli, G., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the twelfth international conference of the CLEF Association. LNCS (12880). Springer https://doi.org/10.1007/978-3-030-72240-1_75
93. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830
94. Yosinski J, Clune J, Bengio Y, Lipson H (2014) How transferable are features in deep neural networks?. In: Advances in neural information processing systems. pp 3320–3328
95. Sharif Razavian A, Azizpour H, Sullivan J, Carlsson S (2014) CNN features off-the-shelf: an astounding baseline for recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp 806–813
96. Ozbulak G, Aytar Y, Ekenel HK (2016) How transferable are cnn-based features for age and gender classification? In: Proceedings of the international conference of the biometrics special interest group. pp 1–6. https://doi.org/10.1109/BIOSIG.2016.7736925
97. Oquab M, Bottou L, Laptev I, Sivic J (2014) Learning and transferring mid-level image representations using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. CVPR '14. IEEE, pp 1717–1724
98. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
99. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. CVPR '17. IEEE, pp 4700–4708

100. Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K (2016) Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5 mb model size. arXiv:1602.07360
101. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861
102. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: Proceedings of the ieee conference on computer vision and pattern recognition. cvpr '09. IEEE, pp 248–255
103. Kingma DP, Ba J (2015) Adam: a method for stochastic optimization. In: Proceedings of the International conference on learning representations