CrisisViT: A Robust Vision Transformer for Crisis Image Classification

Zijun Long
University of Glasgow
z.long.2@research.gla.ac.uk

Richard McCreadie
University of Glasgow
richard.mccreadie@glasgow.ac.uk

Muhammad Imran
Qatar Computing Research Institute
Hamad Bin Khalifa University
mimran@hbku.edu.qa

ABSTRACT

In times of emergency, crisis response agencies need to quickly and accurately assess the situation on the ground in order to deploy relevant services and resources. However, authorities often have to make decisions based on limited information, as data on affected regions can be scarce until local response services can provide first-hand reports. Fortunately, the widespread availability of smartphones with high-quality cameras has made citizen journalism through social media a valuable source of information for crisis responders. However, analyzing the large volume of images posted by citizens requires more time and effort than is typically available. To address this issue, this paper proposes the use of state-of-the-art deep neural models for automatic image classification/tagging, specifically by adapting transformer-based architectures for crisis image classification (CrisisViT). We leverage the new Incidents1M crisis image dataset to develop a range of new transformer-based image classification models. Through experimentation over the standard Crisis image benchmark dataset, we demonstrate that the CrisisViT models significantly outperform previous approaches in emergency type, image relevance, humanitarian category, and damage severity classification. Additionally, we show that the new Incidents1M dataset can further augment the CrisisViT models resulting in an additional 1.25% absolute accuracy gain.

Keywords

Social Media Classification, Crisis Management, Deep Learning, Vision transformers, Supervised Learning

INTRODUCTION

Crisis events, such as floods, fires and COVID-19, generate significant attention from both news media and the general public, leading to related content being posted to a wide variety of social media platforms, such as Twitter or Facebook. Previous studies (Kumar et al. 2011; Dosovitskiy et al. 2020; To et al. 2017) have demonstrated the importance of using social media as a way to acquire information during a crisis event. However, the limited time available to emergency responders in combination with the large volume of posts made on these platforms necessitates automated tooling to extract only the actionable portions of that content (McCreadie et al. 2020; Widener and W. Li 2014). Indeed, over the last decade there have been a wide range of works examining how machine learning can be used to aid emergency responders in finding useful information during crises, primarily focused on analysing the text of posted messages (He, Zhang, et al. 2016; Devlin et al. 2019; Gao et al. 2019).

However, more recently there has been growing interest in the value-add of posted photos and other imagery during an emergency (Imran, Castillo, Diaz, et al. 2015; Said et al. 2019; Buntain et al. 2022). Some papers aim to improve effectiveness of image classification on crisis imagery contents (Imran, Alam, et al. 2020; Asami et al. 2022; X. Li, Caragea, et al. 2019; X. Li and Caragea 2020). Indeed, some studies have shown that images posted on social media

for events such as floods or wildfires can be useful when allocating resources or estimating damage severity (Nguyen, Ofli, et al. 2017; Mouzannar et al. 2018; Sosea et al. 2021; Akhtar et al. 2021). As a result, the development of automated tooling to analyse crisis imagery and categorize it into useful types is of growing importance. To-date, deep convolutional neural networks, e.g., ResNet and VGG16 (Nguyen, Joty, et al. 2016), have been a popular solution for crisis image content categorization, that have reported high accuracy. These solutions rely on general pre-trained models produced from non-crisis image datasets as a starting point, e.g. ImageNet, and then fine-tune those models for a downstream task via transfer learning. It is not obvious why a model initially trained to identify mundane objects like cats or buildings would be effective for identifying images of people needing to be rescued. Moreover, recent advances in the field of computer vision have introduced alternative transformer-based neural architectures (Dosovitskiy et al. 2020), which are suitable for large-scale multi-task pre-training.

Hence, in this paper, we examine whether we can improve the performance of crisis image classification tasks via models pre-trained using in-domain crisis imagery, rather than relying on a general image classification model as a starting point. In particular, using the state-of-the-art ViT architecture (Dosovitskiy et al. 2020) as a base we pre-train new models using crisis imagery from the new incidents1M image dataset (Weber et al. 2022), which we refer to as CrisisViT models. We have released these models for the community, and they can be downloaded via:

- https://github.com/longkukuhi/CrisisViT

Through experimentation over the Crisis Image Benchmark dataset (Nguyen, Joty, et al. 2016), we show that CrisisViT is more effective than previous state-of-the-art deep convolutional neural models, with an increase in accuracy of 3.90 absolute points (from 79.18% to 83.07%). Moreover, the proposed best CrisisViT outperforms all baselines, as well as an existing ViT model, by up to 1.25% absolute accuracy on average. This demonstrates that a dedicated large-scale crisis image dataset is key for the crisis image content categorization task.

RELATED WORK

Image content from Social Media platforms for Crisis Response: Social media is increasingly seen as a critical information and communication platform during emergencies, as a channel to gather and analyze urgent information during a crisis (To et al. 2017; Yin et al. 2015; Shekhar and Setty 2015). However, the majority of prior work in this space has focused on analysing textual content posted to these platforms rather than imagery (Imran, Castillo, Diaz, et al. 2015). On the other hand, recent works have begun to explore the value-add that crisis images posted to social media platforms can bring, as well as how to minimise the costs associated to image analysis through AI automation. For example, Nguyen, Ofli, et al. 2017 demonstrated that crisis images on social media can be used for a variety of humanitarian aid activities (such as identifying areas in need of goods and services). Meanwhile, Alam, Imran, et al. 2017 showed that social media images are helpful for damage assessment during flooding events, while Daly and Thom 2016 illustrated that geotagged images can be used to identify affected regions in need of aid. Functionally, crisis image analysis can be seen as a classification or tagging problem, where a human or machine needs to analyse the image and then assign a label or labels to that image. To-date the crisis image domain has largely focused on four image classification use-cases:

- *Disaster Type Detection*: The high-level classification of the type of disaster depicted within an image, such as an earthquake, fire or flood.

- *Informativeness/Usefulness*: The classification of images to determine whether it contains some form of valuable information for an emergency responder. Typically represented as a binary informative/not informative classification.

- *Humanitarian Categories*: This form of classification is focused on what is happening within the image, where the goal is to identify images that are relevant to different types of humanitarian response activities. Common humanitarian categories include images of affected individuals, images of infrastructure or utility damage, or images of people needing rescue.

- *Damage Severity*: Finally, one common use for crisis images is to judge the severity of damage in a particular area, which is useful for response prioritization or damage costing purposes. The damage severity task mainly targets three levels: severe damage, mild damage, and little or no damage.

Notably, Nguyen, Joty, et al. 2016 developed a standard dataset that combines training and test examples for all four tasks, which we use later in this paper to evaluate our models.

Deep Neural Networks for Image Classification: In the wider field of image classification, the most dominant type of solution is the deep learned AI model. These approaches function by taking an embedding of the pixel data from the image as input, which is fed into a deep neural network that extracts some meaning from that image. A deep neural model is trained by example, where an image is provided to the model, the model then generates a predicted label, and then depending on whether it got the label correct feedback is transferred backward into the model, updating the network. Over the course of seeing thousands to millions of example images, the model learns what pixel embeddings correlate with the desired labels. However, deep neural networks are computationally expensive to train, and tend to exhibit higher accuracy if pre-trained on multiple related tasks (Girshick et al. 2014; Ren et al. 2017; Redmon et al. 2016). Hence it is common for companies and researchers to release pre-trained neural models, which other developers can then adapt at a lower cost to their own use-case (known as transfer learning). With regard to the structure of the neural model itself, there are currently two competing architectures: convolutional neural networks (CNNs); and transformers. CNNs have traditionally been the dominant neural network type used for image classification. Transformer architectures have been shown to be highly effective for text classification (Devlin et al. 2019), but under-perform when adapted for images due to the markedly higher dimensionality (there are more pixels in an image than words in a sentence). Architecturally CNNs are advantaged here, as their convolutional structure forces them to find the parts of the image that matter and discard the rest, enabling them to better generalize to unseen examples. As a result, pre-trained CNNs are popular choices as baselines, such as ResNet152 (He, Zhang, et al. 2016) and VGG (Simonyan and Zisserman 2015).

Advances in Image Transformers: Over the past 5 years, significant research efforts have been made to improve the effectiveness of transformer architectures (Vaswani et al. 2017) for image classification. The issue with applying transformers for image classification is two-fold: 1) training transformers on images is much more costly in comparison to a CNN, as transformer training time rapidly scales with input dimensionality due to its attention mechanism; and 2) transformers when applied to images have been shown to not generalize well to unseen images, as they lack some of the inductive biases that are learned naturally by CNNs. Early approaches tried to reduce training costs by applying self-attention to only the local neighbourhood for each query pixel (Parmar et al. 2018), or by applying attention to only small parts of the image (Weissenborn et al. 2020). However, an important breakthrough occurred with the development of the ViT model (Dosovitskiy et al. 2020), which was the first vision transformer model to efficiently apply attention globally with minimal modifications to the transformer architecture. In 2021, Masked Autoencoders (MAE) (Kumar et al. 2011) were proposed, which then further addressed the cost of training via the use of a high image masking strategy with an encoder-decoder self-supervised pre-training schema, which enables MAE to learn how to reconstruct the original image based on only partial observations of that image. This approach reduces the number of pixels that need to be fed into the transformer and is the best current solution for reducing training time. Similarly, SimMIM (Xie et al. 2022) proposed the use of masked image modelling to pre-trained vision transformers but without a decoder. Meanwhile, Data2vec (Baevski et al. 2022) introduced a teach-student mode to pretrain vision transformers by representation learning. These models are normally pre-trained based on large-scale image datasets like ImageNet Russakovsky et al. 2015, and they can then be 'fine-tuned' with new examples to transfer the pre-trained knowledge into the target downstream tasks, a process that is referred to as transfer learning Torrey and Shavlik 2010. Although vision transformers dominate in the computer vision domain, the transferability of vision transformers remain unclear. As pointed out in Dosovitskiy et al. 2020, lacking discernible learned inductive biases limits the performance of vision transformers to handle downstream tasks. This appears to be a core limitation of transformers that cannot be easily overcome, leading to works such as Zhou et al. 2021 performing expensive whole network fine-tuning (that requires a large in-domain training dataset) to adapt the pre-trained model. In this work, we aim to push the boundaries of crisis image classification performance using transformers, and hence we need such a large in-domain crisis dataset. In our subsequent experiments, we evaluate whether the newly released incidents1M dataset (Weber et al. 2022) is sufficient to enable effective vision transformer models for crisis image classification.

METHODOLOGY

In this work, we investigate whether pre-training on a large-scale crisis image dataset can improve the performance of crisis classification tasks. We choose a state-of-the-art transformer-based image classification model, ViT (Dosovitskiy et al. 2020), as our base model and propose a new CrisisViT variant, which surpasses other deep learning image models in performance and robustness for a range of crisis image classification tasks. We use incidents1M (Weber et al. 2022) as the large-scale crisis image dataset for training. We try out various ways to pretrain CrisisViT, such as different pre-training strategies, examples used, and training labels, based on the dataset characteristics. We discuss the implementation of these models below. When building the CrisisViT models, there are two main decisions that need to be made: 1) determine the dataset used to train; and 2) decide how to train with that dataset.

Pretrain datasets: For the pre-training dataset, we have the option of using either a known effective general image classification dataset, or attempt training with a more specialised in-domain dataset. To represent a general image classification dataset we use the popular ImageNet-1k image collection, while for an in-domain dataset we experiment with the new incidents1M crisis image collection:

- **ImageNet-1k**: The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Deng et al. 2009) project provides more than 14 million human-annotated images that could be used to train deep neural image models. These images are labelled based on the objects depicted, which might be mundane objects like cell phones, animals, or structures. Using such a dataset to pretrain a neural image classification model directs it towards the presence of known objects when later used for a downstream classification task. For example, it is intuitive that if trying to identify images about wildfires, being able to identify a firetruck in the image would be useful. In order to compare our result with other works, we follow the settings of Nguyen, Joty, et al. 2016 and He, Chen, et al. 2022 that use the ImageNet-1k subset of ILSVRC, which is its most commonly used component. ImageNet-1k has 1,000 object classes (types of objects) and contains 1,281,167 training images, 50,000 validation images and 100,000 test images. In the rest of the paper, we refer to this dataset as ImageNet-1k.

- **Incidents1M**: Incidents1M is a large-scale crisis dataset of images taken during natural disasters. This dataset contains annotated labels for 1,787,154 images with two main types: 43 incident categories (e.g. airplane accident, bicycle accident, car accident.); and 49 place categories (e.g. building outdoor, highway, forest, ocean, sky.). Of the 1,787,154 images, just over half (977,088) contain one positive label, i.e. they belong to at least one of the (43+49) categories. It is possible for images to have multiple labels (belong to multiple categories). Notably, Weber et al. 2022 does not release the image files, instead providing URLs pointing to those image files for other researchers to download. As such, over time as online content gets deleted or becomes unavailable this dataset will shrink. We downloaded this dataset during the final quarter of 2022, and at that time we retrieved 1,226,943 of the images (68.7%). This crawled subset has 671,506 images labeled positively to one or more of the incident type categories and 522,782 images labeled positively to one or more of the place categories.

Meanwhile, for the ways of utilising mentioned pre-train datasets, we can use these datasets either in isolation or together:

- **ImageNet-1k + Incidents1M**: Under this setting we load the pre-trained weights from MAE (He, Chen, et al. 2022)—a model prebuilt via self-supervised learning on ImageNet-1k—to perform supervised pre-training on the ImageNet-1k object labels to encode information about the thousand object classes into the model. We further augment this model using training examples from the Incidents1M dataset to teach the model how to identify crisis-related information. This forms a new base model that we can later be fine-tuned for different (crisis image classification) downstream tasks.

- **Incidents1M only**: In this setting, instead of starting from an existing model, we take a blank model and conduct both self-supervised and supervised training using the images and labels in the Incidents1M dataset. This should act similarly to the above base model, but will lack some of the more general object recognition capabilities. In our later experiments, we compare these base models to determine whether starting from a more general image classification model or using only in-domain training is sufficient.

Pretrain tasks: Importantly, the Incidents1M dataset (Weber et al. 2022) supports two main crisis categorization tasks: 1) incident type classification with 43 incident types; and 2) place type classification with 49 place types. In effect, this means that we can pre-train our base model using some or all of these 43+49 image types. We experiment with four ways to utilise these training images and labels in our experiments:

- **Binary training**: As we have 43+49 labeled image types, one methodology for pre-training a base model would be to consider each of these 92 image types as a different binary classification task. We can then incrementally train the base model to classify each of these 92 types in sequence, thereby incrementally building up the model's ability to identify different types of crisis content. We refer to this as Binary classification pre-training and use it as a baseline. However, we remark that this might not be the best training strategy, as lessons learned by the model when training during types seen early may be overwritten by later types (a phenomenon known as catastrophic forgetting).

- **Incident OR Places training**: The second training methodology that we might employ would be to instead combine the images and labels for only one of the Incidents1M tasks, i.e. the 43 incident categories or the 49 place categories, into a unified set of training examples. In this way, we can see which of the two Incidents1M tasks provides more useful information for enhancing our downstream tasks. In contrast to binary training, in this setting, we do not divide our training by image type and instead train all image types for our selected task concurrently. Also, since an image can have multiple labels, we want to avoid sharing images across categories, in this scenario if an image has multiple labels, we consider it as belonging to only the category denoted by the first listed label.

- **Dual (Incident+Places) training**: The final pre-training methodology we use is duel training, which is identical to Incident or Places training, with the exception that we combine both tasks, rather than building separate models for each of the two Incidents1M tasks.

- **Self-supervised training**: Following the self-supervised training method from Masked Autoencoders (Kumar et al. 2011), the CrisisViT model is trained using a self-supervised approach in which it learns to predict missing patches of an image by masking out random portions of the input image and then reconstructing the masked image. By employing this technique, the CrisisViT model can extract meaningful representations from large amounts of unlabeled data, leading to improved performance on image classification tasks.

All variants of CrisisViT use the same architecture as the ViT-base model (Dosovitskiy et al. 2020), but with different hyperparameters. If the pertaining datasets are ImageNet-1k plus Incidents1M, it means we load the pretrain weights from (Dosovitskiy et al. 2020).

EXPERIMENTAL SETUP

Downstream (Target) Dataset: To evaluate how effective the CrisisViT model variants are, we require a downstream or target dataset, which represents one or more meaningful crisis image classification tasks. As discussed previously, the most common uses for crisis imagery are disaster type detection, informativeness/usefulness classification, grouping images into humanitarian categories, and damage severity estimation. Crisis Image Benchmark (Nguyen, Joty, et al. 2016) is a composite test collection that aggregates several datasets together, including CrisisMMD (Alam, Ofli, et al. 2018), data from AIDR (Imran, Castillo, Lucas, et al. 2014) and the Damage Multimodal Dataset (DMD) (Mouzannar et al. 2018). The dataset consists of labels for four tasks:

- Task 1: Disaster type classification

    - Earthquake

    - Fire

    - Flood

    - Hurricane

    - Landslide

    - Other disaster type

    - Not disaster

- Task 2: Informativeness

    - Informative

    - Not informative

- Task 3: Classification into humanitarian categories

    - Affected, injured, or dead people

    - Infrastructure and utility damage

    - Rescue volunteering or donation effort

    - Not humanitarian

- Task 4: Classification into damage severity categories

    - Severe damage

Model	Type	Self-Supervised Dataset	Supervised Dataset	Supervised Methodology	Epochs	Disaster Test	Info Test	Human Test	Damage Test	AVG	Training Time (hours)
ResNet101	CNN	None	ImageNet-1k	Multi-Class (1k)	10	81.3	85.2	76.5	73.7	79.175	N/A
EffiNet (b1)	CNN	None	ImageNet-1k	Multi-Class (1k)	10	81.6	86.3	76.5	75.8	80.05	N/A
VGG16	CNN	None	ImageNet-1k	Multi-Class (1k)	10	79.8	85.8	77.3	75.3	79.55	N/A
ViT-Base	TF	ImageNet-1k	ImageNet-1k	Multi-Class (1k)	20	84.10	86.59	79.43	77.18	81.82	N/A
CrisisViT	TF	Incidents1M	Incidents1M	Multi-Class (Incident)	10	84.91	86.85	79.43	77.96	82.29	36
CrisisViT	TF	Incidents1M	Incidents1M	Multi-Class (Incident)	20	84.73	86.61	79.60	77.31	82.06	420
CrisisViT	TF	Incidents1M	Incidents1M	Multi-Class (Places)	10	84.95	87.85	80.16	**78.75**	82.93 *	420
CrisisViT	TF	Incidents1M	Incidents1M	Multi-Class (Places)	20	**85.26**	**87.97**	80.34	78.72	**83.07** *	420
CrisisViT	TF	Incidents1M	Incidents1M	Multi-Class (Incident+Places)	10	85.01	86.85	79.60	77.31	82.19	430
CrisisViT	TF	Incidents1M	Incidents1M	Multi-Class (Incident+Places)	20	84.57	86.69	79.23	77.41	81.98	430
CrisisViT	TF	ImageNet-1k	ImageNet-1k+Incidents1M	Multi-Class (1k) + Multi-Class (Incident)	10	84.88	87.17	79.64	78.29	82.49	34
CrisisViT	TF	ImageNet-1k	ImageNet-1k+Incidents1M	Multi-Class (1k) + Multi-Class (Incident)	20	85.23	87.08	79.71	78.47	82.62 *	48
CrisisViT	TF	ImageNet-1k	ImageNet-1k+Incidents1M	Multi-Class (1k) + Multi-Class (Places)	10	85.04	87.04	80.15	78.16	82.60 *	34
CrisisViT	TF	ImageNet-1k	ImageNet-1k+Incidents1M	Multi-Class (1k) + Multi-Class (Places)	20	85.04	87.51	79.88	78.01	82.61 *	48
CrisisViT	TF	ImageNet-1k	ImageNet-1k+Incidents1M	Multi-Class (1k) + Binary (Incident+Places)	20	81.13	84.15	75.60	75.71	79.14	460
CrisisViT	TF	ImageNet-1k	ImageNet-1k+Incidents1M	Multi-Class (1k) + Multi-Class (Incident+Places)	10	85.01	87.13	**80.42**	78.19	82.69 *	36
CrisisViT	TF	ImageNet-1k	ImageNet-1k+Incidents1M	Multi-Class (1k) + Multi-Class (Incident+Places)	20	84.95	86.92	79.40	77.70	82.24	52

Table 1. Experimental result overall baselines and variants of CrisisViT model. We use ∗ to denote a significant difference between the performances of the ViT baseline and the proposed model, according to the paired t-test with the Holm-Bonferroni correction for p < 0.01.

- – Mild damage

- – Little or none

This crisis image benchmark provides both training and testing for the four tasks. We follow the same experimental setup for the training, validation, and testing splits as in the original paper (Nguyen, Joty, et al. 2016).

Metrics: We evaluate the performance of CrisisViT models in terms of their classification accuracy. Note that all metrics are reported on the same test set of the Crisis Image Benchmark dataset. Each experiment is run at least three times, and we report the average of the results.

Baselines: Our overall goal in this paper is to determine to what extent a large-scale crisis image dataset (IncidentM1 in this case) improves the performance of transformer-based image classification models when performing crisis content categorization, as well as to determine best practices during training. Hence, in our later experiments we will compare our CrisisViT model to a number of either popular or state-of-the-art image classification models fine-tuned and evaluated on the crisis image benchmark, but that do not have knowledge on the Incidents1M dataset:

- **ResNet101**: (He, Zhang, et al. 2016) proposes ResNet, a convolutional neural network with a deep residual connection, which achieves very high accuracy on the ImageNet dataset. ResNet101 is a deeper variant of ResNet with 101 layers.

- **EffiNet (b1)**: (Tan and Le 2019) studies model scaling and identifies that carefully balancing network depth, width, and resolution can lead to better performance. EffiNet (b1) is the second smallest version of EffiNet, which achieves similar performance to ResNet101 but with a smaller size.

- **VGG16**: A convolutional neural network proposed by (Simonyan and Zisserman 2015) that performs well on a wide range of tasks. It has 16 convolutional layers.

- **ViT-Base**: ViT (Dosovitskiy et al. 2020) is the first vision transformer model to efficiently apply attention globally with minimal modifications to the transformer architecture, achieving remarkable performance on various datasets. ViT-Base is the base version of ViT with 12 layers and 768 hidden size dimension.

CrisisViT Parameters: As with all machine learning models, there are a number of hyper-parameters that can affect the performance of the resultant model. We pre-trained CrisisViT with self-supervised learning and supervised learning on the Incidents1M dataset by using the Adam optimiser, a batch size of 1024 and 128, respectively, and the ReLU activation function. We also experimented with other batch sizes [32,64,128,256,512], which led to lower performance of ViT. We fixed the training epoch for self-supervised training on the Incidents1M dataset at 400, and separately tested the models with supervised training pre-training steps of 10 and 20 epochs on the same dataset. The performances for each experiment are reported separately.

EXPERIMENTAL RESULTS

To evaluate what extent the Incidents1M large-scale crisis image dataset can increase the performance of the image classification models for a range of tasks, we divide our analysis into three primary research questions, based on the different ways that Incidents1M can be utilised:

- RQ1: To what extent can transformer-based architectures outperform traditional convolutional neural networks (CNNs) in image classification tasks?

- RQ2: What are the optimal pre-training strategies for the CrisisViT model when pre-training on the Incidents1M dataset?

- RQ3: Does starting from a more general image classification model, such as ImageNet-1k, provide significant advantages over using only in-domain training with the Incidents1M dataset in terms of the performance and robustness of the CrisisViT model for crisis image classification tasks?

In this section, we report the results comparing the performances between state-of-the-art deep convolutional neural image models, the transformer-based image model ViT, and different variants of our proposed CrisisViT model, produced for crisis content categorization on four tasks of the Crisis image benchmark dataset.

RQ1: ViT vs. Convolutional neural baselines

We begin by determining how well a transformer image classification architecture like ViT performs for the domain of crisis image classification. Since most prior works (as discussed in the related work) employed the convolutional neural network (CNN) architecture, we intend to understand if transformer-based models make a difference. To this end, we compare three CNN baselines, namely ResNet101, EffiNet (b1) and VGG16, with the best transformer architecture ViT. What differentiates the three cases is their training starting point, i.e., the base model. We train the base model for four downstream (target) tasks to produce corresponding four models. The first four rows of Table 1 report the performance of these models under the test set for each task.

As shown in Table 1, Vit outperforms the other three CNN-based models. Specifically, the ViT transformer model appears to be primarily advantaged when used for Disaster Type classification, Humanitarian category classification, and Damage Severity estimation, with reported gains over the next best CNN-based model by 3.5%, 1.4% and 1.4% absolute classification accuracy, respectively. Meanwhile, a much smaller but notable gain of 0.3% is observed for the Informativeness classification. Overall, this confirms our expectation that transformer-based models are the current state-of-the-art in this domain, and as such, we use ViT as our primary comparison point in the remainder of this paper.

RQ2: Pre-training using Incident Types and Place Categories

Having shown that the Transformer architecture ViT is superior to prior CNN-based architectures and quantified our baseline performance, we now examine the core question of this work: can we further boost the performance using a large-scale crisis image dataset? The underlying rationale is that by teaching a deep neural model how to tackle a different but related task helps the model learn the downstream task better. In this scenario, we used the Incidents1M dataset with labels for 43 incident-type and 49 place categories. These tasks are different but conceptually related to our four downstream tasks (i.e., disaster type classification, informativeness/Usefulness, humanitarian category classification, and damage severity estimation). Rows 5-10 of Table 1 report the accuracy of the ViT architecture when pre-trained using Incidents1M instead of ImageNet-1k (as was used in the ViT-Base baseline). We report performance when pre-training with Incident type labels (Multi-Class (Incident)), place categorization labels (Multi-Class (Place)), and both (Multi-Class (Incident+Places)). We also report performances for 10 and 20 epochs to illustrate how performance improves with more training time.

From Table 1, we observe that in nearly all cases pre-training on Incidents1M leads to superior performances for crisis image classification than using only ImageNet-1k. For instance, when pre-training with the Incidents1M place category labels for 20 epochs, we observe a statistically significant ($p \leq 0.05$) accuracy gain over ViT-Base of 1.16%. Second, comparing the Incident and Place labels provided by Incidents1M, the Place labels result in the best-performing downstream models in all cases, while the Incident labels provide a comparably smaller benefit (and in one case it harms the performance). Furthermore, we notice that when combining the Incident and Place labels together, performance does not improve over using the Place labels alone, indicating that the Incident labels are redundant when the Place labels are available. Third, comparing the effect of providing more training, when

pre-training using the place labels, more training time (20 epochs) does lead to small performance gains (around 0.1-0.3%). However, when providing additional training time to the Incident labels, downstream accuracy tends to degrade, confirming that teaching the model about the Incidents1M incident types leads to questionable benefits. Overall, we can conclude that pre-training with an in-domain dataset can lead to performance gains. Indeed, we observed between 0.9% and 1.5% accuracy across the downstream tasks tested. On the other hand, the inconsistent performance of the models pre-trained with the Incident labels indicates that not all conceptually related tasks are useful evidence for pre-training, and so researchers and developers should be selective regarding what datasets to use for pre-training.

RQ3: ImageNet-1k + Incidents1M?

In the previous set of experiments we replaced ImageNet-1k-based pre-training with pre-training using the in-domain dataset i.e., Incidents1M. However, given that ImageNet-1k forms the basis for many effective image classification models in the literature, as well as providing a strong baseline (via ViT-Base), it is worth investigating whether we can instead augment ImageNet-1k training instead of replacing it. Hence, we pre-train a second set of base models that take the same base model as ViT-Base subject to ImageNet-1k self-supervised and supervised training and then add further pre-training using the Incidents1M data. If ImageNet-1k provides value on top of Incidents1M, combining both should result in a small improvement in accuracy on the downstream tasks. The final seven rows of Table 1 report classification accuracy on the four downstream tasks when we combine ImageNet-1k and Incidents1M pre-training.

In Table 1 we can observe that the best performing Incidents1M models pre-trained on the Places labels does not show performance improvements despite providing additional ImageNet-1k training. This might lead us to conclude that ImageNet-1k is not adding value. However, if we investigate both the Incidents1M models that are pre-trained on the Incident or Incident+Places labels, we do observe a small performance uplift for the majority of tasks. Indeed, one of these models achieved the best overall performance for the Humanitarian categorisation task. On the other hand, given the small degree of the performance difference, it is not apparent that starting from the pre-trained ViT-Base model is better than training a new model.

CONCLUSIONS

Social media has become an increasingly important platform for emergency response agencies to obtain valuable information, particularly images, for various crisis response tasks. However, due to the sheer volume of content on social media, automated techniques for filtering and classifying images are necessary. Existing methods rely on convolutional neural networks pre-trained on the general ImageNet-1k dataset, but recent developments in transformer-based image classifiers in combination with increased availability of tagged crisis imagery (via the Incidents1M dataset) have opened up new possibilities. In this paper, we introduced CrisisViT, a transformer-based architecture pre-trained on the Incidents1M dataset, which can be adapted for multiple downstream crisis image classification tasks. Through experimentation on the four tasks (Disaster Type classification, Informativeness classification, Humanitarian Category classification, and Damage Severity estimation) supported by the Crisis Image Benchmark dataset, we show that pre-training on the Incidents1M dataset can lead to significant improvements in accuracy, with an average absolute gain of 1.25% over the four crisis image classification tasks tested. This work represents an important step towards building more effective crisis response tools that can utilize social media image data to support emergency response efforts.

REFERENCES

Akhtar, Z., Ofli, F., and Imran, M. (2021). "Towards Using Remote Sensing and Social Media Data for Flood Mapping". In: *18th International Conference on Information Systems for Crisis Response and Management, ISCRAM 2021, Blacksburg, VA, USA, May 2021*. Ed. by A. Adrot, R. Grace, K. A. Moore, and C. W. Zobel. ISCRAM Digital Library, pp. 536–551.

Alam, F., Imran, M., and 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 2017*, pp. 601–604.

Alam, F., Ofli, F., and Imran, M. (2018). "CrisisMMD: Multimodal Twitter Datasets from Natural Disasters". In: *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018*. AAAI Press, pp. 465–473.

Asami, K., Fujita, S., Hiroi, K., and Hatayama, M. (2022). "Data Augmentation with Synthesized Damaged Roof Images Generated by GAN". In: *19th International Conference on Information Systems for Crisis Response and Management, ISCRAM 2022, Tarbes, France, May 22-25, 2022*. Ed. by R. Grace and H. Baharmand. ISCRAM Digital Library, pp. 256–265.

Baevski, A., Hsu, W., Xu, Q., Babu, A., Gu, J., and Auli, M. (2022). "data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language". In: *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*. Ed. by K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato. Vol. 162. Proceedings of Machine Learning Research. PMLR, pp. 1298–1312.

Buntain, C., McCreadie, R., and Soboroff, I. (2022). "Incident Streams 2021 Off the Deep End: Deeper Annotations and Evaluations in Twitter". In: *19th International Conference on Information Systems for Crisis Response and Management, ISCRAM 2022, Tarbes, France, May 22-25, 2022*. Ed. by R. Grace and H. Baharmand. ISCRAM Digital Library, pp. 584–604.

Daly, S. and Thom, J. A. (2016). "Mining and Classifying Image Posts on Social Media to Analyse Fires". In: *13th Proceedings of the International Conference on Information Systems for Crisis Response and Management, Rio de Janeiro, Brasil, May 22-25, 2016*. Ed. by A. H. Tapia, P. Antunes, V. A. Bañuls, K. A. Moore, and J. P. de Albuquerque. ISCRAM Association.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "ImageNet: A Large-Scale Hierarchical Image Database". In: *CVPR09*.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Ed. by J. Burstein, C. Doran, and T. Solorio. Association for Computational Linguistics, pp. 4171–4186.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., and Gelly, S. (2020). "An image is worth 16x16 words: Transformers for image recognition at scale". In: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Gao, P., Jiang, Z., You, H., Lu, P., Hoi, S. C. H., Wang, X., and Li, H. (2019). "Dynamic Fusion With Intra- and Inter-Modality Attention Flow for Visual Question Answering". In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, pp. 6639–6648.

Girshick, R. B., Donahue, J., Darrell, T., and Malik, J. (2014). "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. IEEE Computer Society, pp. 580–587.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. B. (2022). "Masked Autoencoders Are Scalable Vision Learners". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, pp. 15979–15988.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, pp. 770–778.

Imran, M., Alam, F., Qazi, U., Peterson, S., and Ofli, F. (2020). "Rapid Damage Assessment Using Social Media Images by Combining Human and Machine Intelligence". In: *17th International Conference on Information Systems for Crisis Response and Management, ISCRAM 2020, May 2020*. Ed. by A. L. Hughes, F. McNeill, and C. W. Zobel. ISCRAM Digital Library, pp. 761–773.

Imran, M., Castillo, C., Diaz, F., and Vieweg, S. (2015). "Processing Social Media Messages in Mass Emergency: A Survey". In: *ACM Comput. Surv.* 47.4, 67:1–67:38.

Imran, M., Castillo, C., Lucas, J., Meier, P., and Vieweg, S. (2014). "AIDR: Artificial intelligence for disaster response". In: *Proceedings of WWW*. ACM.

Kumar, S., Barbier, G., Abbasi, M. A., and Liu, H. (2011). "TweetTracker: An Analysis Tool for Humanitarian and Disaster Relief". In: *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*. The AAAI Press.

Li, X. and Caragea, D. (2020). "Improving Disaster-related Tweet Classification with a Multimodal Approach". In: *17th International Conference on Information Systems for Crisis Response and Management, ISCRAM 2020, May 2020*. Ed. by A. L. Hughes, F. McNeill, and C. W. Zobel. ISCRAM Digital Library, pp. 893–902.

Li, X., Caragea, D., Caragea, C., Imran, M., and Ofli, F. (2019). "Identifying Disaster Damage Images Using a Domain Adaptation Approach". In: *Proceedings of the 16th International Conference on Information Systems for Crisis Response and Management, València, Spain, May 19-22, 2019*. Ed. by Z. Franco, J. J. González, and J. H. Canós. ISCRAM Association.

McCreadie, R., Buntain, C., and Soboroff, I. (2020). "Incident Streams 2019: Actionable Insights and How to Find Them". In: *17th International Conference on Information Systems for Crisis Response and Management, ISCRAM 2020, May 2020*. Ed. by A. L. Hughes, F. McNeill, and C. W. Zobel. ISCRAM Digital Library, pp. 744–760.

Mouzannar, H., Rizk, Y., and Awad, M. (2018). "Damage Identification in Social Media Posts using Multimodal Deep Learning". In: *Proceedings of the 15th International Conference on Information Systems for Crisis Response and Management, Rochester, NY, USA, May 20-23, 2018*. Ed. by K. Boersma and B. M. Tomaszewski. ISCRAM Association.

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*, pp. 569–576.

Nguyen, D. T., Joty, S. R., Imran, M., Sajjad, H., and Mitra, P. (2016). "Applications of Online Deep Learning for Crisis Response Using Social Media Information". In: *CoRR* abs/1610.01030. arXiv: `1610.01030`.

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., and Tran, D. (2018). "Image transformer". In: *International Conference on Machine Learning*. PMLR, pp. 4055–4064.

Redmon, J., Divvala, S. K., Girshick, R. B., and Farhadi, A. (2016). "You Only Look Once: Unified, Real-Time Object Detection". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, pp. 779–788.

Ren, S., He, K., Girshick, R. B., and Sun, J. (2017). "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 39.6, pp. 1137–1149.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., and Bernstein, M. (2015). "Imagenet large scale visual recognition challenge". In: *International journal of computer vision* 115.3, pp. 211–252.

Said, N., Ahmad, K., Riegler, M., Pogorelov, K., Hassan, L., Ahmad, N., and Conci, N. (2019). "Natural disasters detection in social media and satellite imagery: a survey". In: *Multimedia Tools and Applications* 78.22, pp. 31267–31302.

Shekhar, H. and Setty, S. (2015). "Disaster analysis through tweets". In: *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, pp. 1719–1723.

Simonyan, K. and Zisserman, A. (2015). "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Y. Bengio and Y. LeCun.

Sosea, T., Sirbu, I., Caragea, C., Caragea, D., and Rebedea, T. (2021). "Using the Image-Text Relationship to Improve Multimodal Disaster Tweet Classification". In: *18th International Conference on Information Systems for Crisis Response and Management, ISCRAM 2021, Blacksburg, VA, USA, May 2021*. Ed. by A. Adrot, R. Grace, K. A. Moore, and C. W. Zobel. ISCRAM Digital Library, pp. 691–704.

Tan, M. and Le, Q. V. (2019). "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks". In: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*. Ed. by K. Chaudhuri and R. Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 6105–6114.

To, H., Agrawal, S., Kim, S. H., and Shahabi, C. (2017). "On identifying disaster-related tweets: Matching-based or learning-based?" In: *2017 IEEE third international conference on multimedia big data (BigMM)*. IEEE, pp. 330–337.

Torrey, L. and Shavlik, J. (2010). "Transfer learning". In: *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI global, pp. 242–264.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Ed. by I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, pp. 5998–6008.

Weber, E., Papadopoulos, D. P., Lapedriza, À., Ofli, F., Imran, M., and Torralba, A. (2022). "Incidents1M: a large-scale dataset of images with natural disasters, damage, and incidents". In: *CoRR* abs/2201.04236. arXiv: 2201.04236.

Weissenborn, D., Täckström, O., and Uszkoreit, J. (2020). "Scaling Autoregressive Video Models". In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Widener, M. J. and Li, W. (2014). "Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US". In: *Applied Geography* 54, pp. 189–197.

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., and Hu, H. (2022). "Simmim: A simple framework for masked image modeling". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9653–9663.

Yin, J., Karimi, S., Lampert, A., Cameron, M. A., Robinson, B., and Power, R. (2015). "Using Social Media to Enhance Emergency Situation Awareness: Extended Abstract". In: *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*. AAAI Press, pp. 4234–4239.

Zhou, H.-Y., Lu, C., Yang, S., and Yu, Y. (2021). "ConvNets vs. Transformers: Whose visual representations are more transferable?" In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2230–2238.