Multimodal Categorization of Crisis Events in Social Media

Mahdi Abavisani*
Department of Electrical and Computer Engineering
Rutgers University
Piscataway, NJ
mahdi.abavisani@rutgers.edu

Liwei Wu*
Department of Statistics
University of California, Davis
Davis, CA
liwu@ucdavis.edu

Shengli Hu
Dataminr Inc., NY, USA
shu@dataminr.com

Joel Tetreault
Dataminr Inc., NY, USA
jtetreault@dataminr.com

Alejandro Jaimes
Dataminr Inc., NY, USA
ajaimes@dataminr.com

Abstract

Recent developments in image classification and natural language processing, coupled with the rapid growth in social media usage, has enabled fundamental advances in detecting breaking events around the world in real-time. Emergency response is one such area that stands to gain from these advances. By processing billions of texts and images a minute, events can be automatically detected to enable emergency response workers to better assess rapidly evolving situations and deploy resources accordingly. To date, most event detection techniques in this area have focused on image-only or text-only only approaches, limiting detection performance and impacting the quality of information delivered to crisis response teams. In this paper, we present a new multimodal fusion method which leverages both images and texts as input and show that it outperforms the singular approaches and strong baselines by a wide margin on three crisis related tasks.

1. Introduction

Each second, billions of images and texts that capture a wide range of events happening around us are uploaded to social media platforms from all over the world. At the same time, the fields of Computer Vision (CV) and Natural Language Processing (NLP) are rapidly advancing [17, 29] and are being deployed at scale. With large-scale visual recognition and textual understanding available as fundamental tools, it is now possible to identify and classify events across the world in real-time. This is possible, to some extent, in images and text separately, and in limited cases, using a combination. A major difficulty in crisis events,[1] in particular, is that as events surface and evolve, users post fragmented, sometimes conflicting information in the form of image-text pairs, making the automatic identification of notable events significantly more challenging.

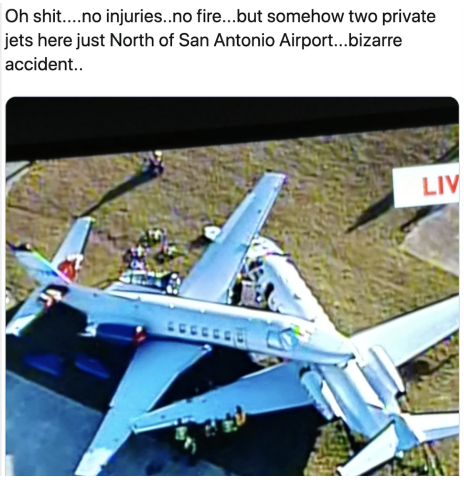

Figure 1. A Crisis-related Image-text Pair from Social Media

Unfortunately, in the middle of a crisis, the information that is valuable for first responders and the general public often comes in the form of image-text pairs. So while traditional CV and NLP methods that treat visual and textual information separately can help, a big gap exists in current approaches. Despite the general consensus on the importance of using AI for social good [27, 23, 2], the power of social media, and a long history of interdisciplinary research on humanitarian crisis efforts, there has been very little work

*Equal contribution, with ordering decided by Python. Research work was done while authors were interning at Dataminr Inc.

[1] An event that is going (or is expected) to lead to an unstable and dangerous situation affecting an individual, group, community, or whole society (Wikipedia); typically requiring an emergency response.

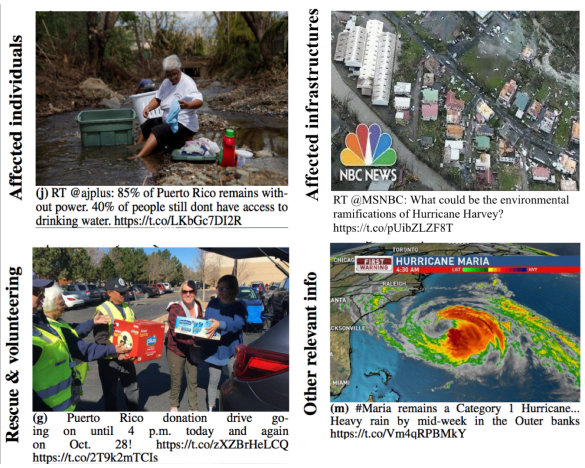

Figure 2. Task 2: Event Classification with Texts and Images

on automatically detecting crisis events *jointly* using visual and textual information.

Prior approaches that tackle the detection of crisis events have focused on either image-only or text-only approaches. As shown in Figure 1, however, an image alone can be ambiguous in terms of its urgency whereas the text alone may lack details.

In this paper, we present a framework to detect crisis events using a combination of image and text information. In particular, we present an approach to automatically label images, text, and image-text pairs based on the following criteria/tasks: 1) **Informativeness**: whether the social media post is useful for providing humanitarian aid in an emergency event, 2) **Event Classification**: identifying the type of emergency (in Figure 1, we illustrate different categories that different image-text pairs belong to in our event classification task), and 3) **Severity**: rating how severe the emergency is based on the damage indicated in the image and text. Our framework consists of several steps in which, given an image-text pair, we create a feature map for the image, word embeddings for the text, and use a cross-modal attention mechanism to fuse information from the two modalities. It differs from previous multi-modal classification in how it deals with fusing that information. Our main contribution can be summarized as follows:

- We present a novel, multimodal framework for processing user-generated content during crises. This approach, "Cross Attention", avoids transferring negative knowledge between modalities and makes use of stochastic shared embeddings [64] to alleviate and prevent overfitting in small data.

We show that this approach directly benefits the three crisis response tasks when compared to image-only and text-only approaches, as well as several strong baselines.

2. Related Work

2.1. AI for Emergency Response

Recent years have seen an explosion in the use of Artificial Intelligence for social good [27, 23, 2]. Social media has proven to be one of most relevant and diverse resources and testbeds, whether it be for identifying risky mental states of users [12, 19, 25], recognizing emergent health hazards [18], filtering for and detecting natural disasters [53, 44, 52], or surfacing violence and aggression in social media [10].

Most prior work on detecting crisis events in social media have focused on text signals. For instance, Kumar *et al*. [39] propose a real-time tweet-tracking system to help first responders gain situational awareness once a disaster happens. Shekhar *et al*. [55] introduce a crisis analysis system to estimate the damage level of properties and the distress level of victims. At a large scale, filtering (e.g., by anomaly or burst detection), identifying (e.g., by clustering), and categorizing (e.g., by classifying) disaster-related texts on social media have been the foci of multiple research groups [59, 61, 67], achieving accuracy levels topping at 0.75 on small annotated datasets collected from Twitter.

Disaster detection in images has been an active front, whether it be user-generated content or satellite images (for a detailed survey, refer to Said *et al*. [53]). For instance, Ahmad *et al*. [3] introduce a pipeline method to effectively link remote sensor data with social media to better assess damage and obtain detailed information about a disaster. Li *et al*. [42] use convolutional neural networks and visualization methods to locate and quantify damage in a disaster images. Nalluru *et al*. [47] combine semantic textual and image features to classify the relevancy of social media posts in emergency situations.

Our framework focuses on combining images and text, yielding performance improvements on image classification tasks in disasters.

2.2. Multimodal Deep Learning

Existing multimodal learning frameworks applied to the crisis domain are relatively limited. Lan *et al*. [40] combine early fusion and late fusion methods to incorporate their advantages, Ilyas [32] introduce a disaster classification system based on naive-bayes classifiers and support vector machines. Kelly *et al*. [36] introduce a system for real-time extraction of information from text and image content in Twitter messages with the spatio-temporal metadata for filtering, visualizing, and thus monitoring flooding events. Mouzannar *et al*. [46] propose a multimodal deep learning framework to identify damage related information on social media posts with texts, images, and video. Nevertheless, none of the above use state-of-the-art models for both vision and

text with attention mechanism.

There is plenty of work in integrating image and text signals with the state-of-the-art architectures and attention mechanisms for other tasks such as image captioning [9], visual question answering [7, 22], and text-image matching [56, 21, 41], as well as multimodal fusion [65, 20, 45, 5]. We categorize both bodies of work by the use of attention mechanisms, as illustrated in Table 1. Multimodal refers to combining more than one modality for a downstream task, whereas cross-modal refers to using one modality for a task that's related to another modality. Among studies of cross-modal or multimodal applications on the left in Table 1 — mostly of image captioning, text-image matching, and visual question answering, there are a few papers close to ours in methodology even though the settings, downstream tasks, and objectives are different. Fukui *et al*. [22], Ilija and Feng [31] apply co-attention between texts and images for visual question answering. While Lee *et al*. [41] and Hessel *et al*. [28] both use cross-attention mechanisms to combine different modalities, theirs differ from ours in that in Lee *et al*. [41] cross-attention was applied for better localization and matching between text and images across layers whereas in Hessel *et al*. [28], cross-attention was introduced after self-attention, which ambiguates its effects.

Table 1. Related Multimodal Work Categorized by Attention Mechanism

Attention	Cross/Multimodal Applications	Multimodal Fusion
None	Frome *et al*. [21], [35]	Ngiam *et al*. [48], Bruni *et al*. [11], [58, 8]
Co-attention	Fukui *et al*. [22], [31]	Lu *et al*. [43], Yang *et al*. [66], [8, 38]
Other	Lee *et al*. [41], Hessel *et al*. [28]	Ours

Multimodal fusion which ranges from model-agnostic [51, 40], to model-based such as Kernel-based [24, 14, 33], graphical model based [26, 34], and neural networks based [63, 48, 15], is an important topic in the field of multimedia analysis. Here, we are specifically interested in model-based fusion methods that enhance well-known neural network architecture for efficient classification. Ngiam *et al*. [48] first use deep autoencoders to learn joint features from audio and video while Srivastava and Salakhutdinov [58] jointly extract visual and textual features with deep Boltzmann machines. Arevalo *et al*. [8], Fukui *et al*. [22] explore complex gating mechanisms and compact bilinear pooling as multi-modal fusion methods. Kiela *et al*. [38] transfer continuous features from neural networks trained on ImageNet classifications using various fusion methods including bilinear, additive, pooling, and gating. Ours differs as we use attention mechanism before concatenation.

The multimodal fusion methods closest to our proposed approach include ViLBERT [43], VisualBERT, LXMERT, and VL-BERT.Existing fusion frameworks are often based on the assumption that all the input modalities contain complimentary information [38, 48, 8]. However, in the application of crisis tweets categorization, one modality may contain non-informative or even misleading information. In these cases, using conventional fusion methods such as Vilbert [43] may degrade performance. Our model does not always fuse the representations from different modalities, but rather mitigates the effects of one modality over another on a case by case basis. The attention module in our model passes information based on the confidence in the usefulness of different modalities. The more confident modality blocks more features from the other modality through their cross-attention link. The partially blocked results of both modalities are later judged by a self-attention layer to decide which information should be passed to the next layer. While our attention module is closely related to co-attention and self-attention mechanisms [62], as apposed to them, it does not need the input features to be homogeneous. Self-attention and co-attention layers can be sensitive to heterogeneous inputs. The details of the model are described in the next section.

3. Methodology

The architecture we propose is designed for classification problems that take as input image-text pairs. Our methodology consists of 4 parts (Figure 3[2]): the first two parts extract feature maps from the image and extract embeddings from the text, respectively; the third part comprises our crossmodal-attention approach to fuse projected image and text embeddings; and the fourth part uses Stochastic Shared Embeddings (SSE) [64] as our regularization technique to prevent over-fitting and deal with training data with inconsistent labels for image and text pairs.

We describe each module in the sub-sections that follow.

3.1. Image Model for Feature Map Extraction

We extract feature maps from images using Convolutional Neural Networks (CNNs). In our model we select DenseNet [30], which reduces module sizes and increases connections between layers to address parameter redundancy and improve accuracy (other approaches, such as EfficientNet [60] could also be used, but DenseNet is efficient and commonly used for this task).

For each image v_i, we therefore have:

$$f_i = \textbf{DenseNet}(v_i), \tag{1}$$

where v_i is the input image, $f_i \in \mathbb{R}^{W \times H \times C}$ is a deep feature map in the DenseNet and W, H, C are the i-th feature map's height, width and number of channels respectively.

3.2. Text Model for Embedding Extraction

Full-network pre-training [49, 17] has led to a series of breakthroughs in language representation learning. Specifi-

[2]the DenseNet and Bert graphs are from Huang *et al*. (2017) and Devlin *et al*. (2018)

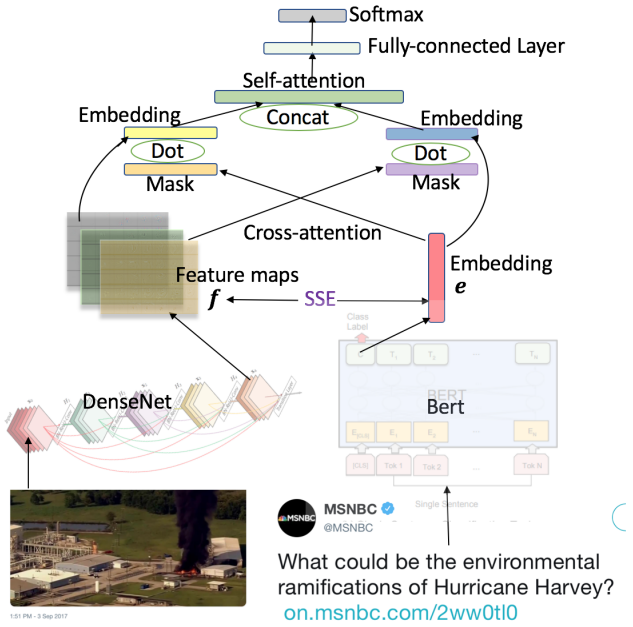

Figure 3. Illustration of Our Framework

cally, deep-bidirectional Transformer models such as BERT [17] and its variants [66, 6] have achieved state-of-the-art results on various natural language processing tasks by leveraging close and next-sentence prediction tasks as weakly-supervised pre-training. Therefore, we use BERT as our core model for extracting embeddings from text (variants such as XLNET and ALBERT could also be used). We use the BERT model pre-trained on Wiki and Books data[3] on crisis-related tweets t_i's. For each text input t_i, we have

$$e_i = \textbf{BERT}(t_i), \tag{2}$$

where t_i is a sequence of word-piece tokens and $e_i \in \mathbb{R}^{756}$ is the sentence embedding. Like in the BERT paper [17], we take the embedding associated with [CLS] to represent the whole sentence.

In the next subsection we detail how the DenseNet and BERT are fused.

3.3. Cross-attention module for avoiding negative knowledge in fusion

After we obtain the image feature map $f_i \in \mathbb{R}^{W \times H \times C}$ (DenseNet [30]) and the sentence embedding $e_i \in \mathbb{R}^D$ (BERT [17]), we use a new cross-attention mechanism to fuse the information they represent. In many text-vision tasks, the input pair can contain noise. In particular, in the tweets classification task, one modality may contain non-informative or even misleading information. In such a case, negative transfer can occur. Our model can mitigate the effects of one modality over another on a case by case basis.

To mitigate this issue, in our cross-attention module, we use a combination of cross-attention layers and a self-attention layer. This module passes information based on the confidence in the usefulness of different modalities. Each modality can block the features of the other modality based on its confidence in the usefulness of its input. This happens with the cross-attention layer. The result of partially blocked features from both modalities is later fed to a self-attention layer to decide which information should be passed to the next layer. We do this, as described next.

We use a fully-connected layer to project the image feature map into a fixed dimensionality K (we use $K = 100$), and similarly project the sentence embedding so that:

$$\tilde{f}_i = F(W_v^T f_i + b_v) \tag{3}$$
$$\tilde{e}_i = F(W_e^T e_i + b_e) \tag{4}$$

where F represents an activation function such as relu (used in our experiments) and both $\tilde{f}_i$ and $\tilde{e}_i$ are of dimension $K = 100$.

In the case of misleading information in one modality, without an attention mechanism (such as co-attention [43]), the resulting $\tilde{f}_i$ and $\tilde{e}_i$ cannot be easily combined without hurting performance. Here, we propose a new attention mechanism called cross-attention (Figure 3), which differs from standard co-attention mechanisms: the attention mask α_{v_i} for the image is completely dependent on the text embedding e_i, while the attention mask α_{e_i} for the text is completely dependent on the image embedding f_i. Mathematically, this can be expressed as follows:

$$\alpha_{v_i} = \sigma(W_v'^T f_i + b_v') \tag{5}$$
$$\alpha_{e_i} = \sigma(W_e'^T e_i + b_e'), \tag{6}$$

where σ is the Sigmoid function. Co-attention, in contrast, can be expressed as follows:

$$\alpha_{v_i} = \sigma(W_v'^T [f_i | e_i] + b_v') \tag{7}$$
$$\alpha_{e_i} = \sigma(W_e'^T [f_i | e_i] + b_e'), \tag{8}$$

where $|$ means concatenation.

After we have the attention masks $\alpha_{v_i}, \alpha_{e_i}$ for image and text respectively, we can augment the projected image and text embeddings $\tilde{f}_i, \tilde{e}_i$ with $\alpha_{v_i} \cdot \tilde{f}_i$ and $\alpha_{e_i} \cdot \tilde{e}_i$ before performing concatenation or adding. In our experiments, we used concatenation, but obtained similar performance using addition.

The last step of this module takes the concatenated embedding which jointly represents the image and text tuple in and feeds into the two-layer fully-connected networks. We add self-attention in the fully-connected networks and use

[3] https://resources.wolframcloud.com/NeuralNetRepository/resources/BERT-Trained-on-BookCorpus-and-English-Wikipedia-Data

the standard softmax cross-entropy loss for this multi-class classification problem.

In Section 4, we show that the combination cross-attention layers and the self attention layer on their concatenation works better than co-attention and self-attention mechanisms for the tasks we address in this paper.

3.4. SSE for Better Regularization

Stochastic Shared Embeddings (SSE) [64] are a data-driven approach to regularizing embedding layers, by stochastically making transitions of embeddings during stochastic gradient descent (SGD). *SSE-Graph*, one of the two SSE variants in [64], requires construction of a knowledge graph over the embeddings. Here, we can treat feature maps of images as embeddings and use class labels to construct knowledge graphs. The feature maps of two images are connected by an edge in the graph, if and only if they belong to the same class for this particular task (e.g. they are both labeled "affected infrastructure").

After constructing the graph, we follow the procedure in the SSE paper [64]: we define transition probability of feature maps from j to k as $p(j, k|\Phi)$, and because it is assumed that when an edge between two feature maps exists, it is more likely that they would be replaced by one another, we set the ratio of $p(j, k|\Phi)$ and $p(j, l|\Phi)$ to be a constant greater than 1. In other words, we have:

$$j \sim k, j \not\sim l \longrightarrow p(j, k|\Phi)/p(j, l|\Phi) = \rho, \tag{9}$$

where ρ is a tuning parameter and $\rho > 1$. We also have:

$$p(j, j|\Phi) = 1 - p_0, \tag{10}$$

where p_0 is called the *SSE probability*.

Both p_0 and ρ are treated as tuning hyper-parameters in experiments and can be tuned fairly easily. With eq. (9), eq. (10) and $\sum_k p(j, k|\Phi) = 1$, we can derive transition probabilities between any two feature maps to fill out the transition probability table.

We do the same for the sentence embeddings obtained by the BERT model. In other words, SSE can be applied to the image modality or to the text modality. In experiments, for an image & sentence pair, we assign equal chances of applying SSE to image side or text side. Dropout regularization [57] is also used for hidden units as usual.

4. Experimental Setup

The image-text classification problem we consider can be formulated as follows: we have as input $(v_1, t_1), \ldots, (v_i, t_i), \ldots, (v_n, t_n)$, where n is the number of training tuples and the i-th tuple consists of both image v_i and text t_i. The respective labels for v_i and t_i's are also given in training data. Our goal is to predict the correct label for any unseen (v, t) pair. To simplify the evaluation, we assume there is only one correct label associated with the unseen (v, t) pairs, so this paper is about tackling a multi-class classification problem instead of a multi-label problem.

4.1. Dataset

There are very few crisis datasets, and to the best of our knowledge there is only one *multimodal* crisis dataset: CrisisMMD [4], which consists of annotated image-tweet pairs where images and tweets are independently labeled as described below. The dataset was collected using event-specific keywords and hashtags during seven natural disasters in 2017 — Hurricane Irma, Hurricane Harvey, Hurricane Maria, the Mexico earthquake, California wildfires, Iraq-Iran earthquakes, and Sri Lanka floods. The corpus is comprised of three types of manual annotations:

Task 1: Informative vs. Not Informative: whether a given tweet text or image is useful for humanitarian aid purposes, defined as providing assistance to people in need;

Task 2: Humanitarian Categories: given an image, or tweet, or a pair of both, categorize it into one of the five following categories:

- Infrastructure, and utility damage;
- Vehicle damage;
- Rescue, volunteering, or donation efforts;
- Affected individuals (injury, dead, missing, found, etc.);
- Other relevant information.

Task 3: Damage Severity: assess the severity of damage reported in a tweet image and classify it into Severe, Mild, and Little/None.

It is important to note that while the annotations for the last task are only on images, our experiments reveal that using tweet texts along with the images can boost performance. In addition, our paper is the first one to perform all three tasks on this dataset (text-only, image-only, combined).

4.2. Settings

Images and text from tweets in this dataset were annotated independently. Thus, in many cases, images and text in the same pairs may not share the same labels for either Task 1 or Task 2 (labels for Task 3 were only created by annotating the images). Given the different evaluation conditions, we carry out three evaluation settings for the sake of being comprehensive in our model assessment but also to establish best practices for the community: *Setting A*: we exclude the image-text pairs with differing labels for image and text; *Setting B*: we include the image-text pairs with dif-

ferent labels in the training set but keep the test set the same as in A. In addition, we introduce *Setting C* to mimic a realistic crisis tweet classification task where we only train on events that have transpired before the event(s) in the test set.

Table 2 shows the number of samples in each set for different setting and tasks.

Setting A: In this setting our train and test data is sampled from tweets in which the text and image pairs have the same label. That is:

$$\mathcal{C}(v_i) = \mathcal{C}(t_i), \tag{11}$$

where $\mathcal{C}(x)$ denotes the class of data point x. This results in a small, yet potentially more reliable training set. We mix the data from all seven crisis events and split the data into training, dev and test sets.

Setting B: We relax the assumption in Equation 11 and allow in training:

$$\mathcal{C}(v_i) \neq \mathcal{C}(t_i), \tag{12}$$

As the training set of this setting contains samples with inconsistent labels for image and text, multimodal fusion methods such as late feature fusion cannot deal with the training data. Our method, on the other hand, with the use of SSE, can exchange the training data with inconsistent label and convert the training pair to a pair with consistent label for both image and text. We do this by manually setting $p_0 = 1$ for the training cases with inconsistent image-text labels. Since unimodal models only receive one of the modalities, it is also possible to train them separately on images and texts and use an average of their prediction in the testing stage (also known as score level fusion).

However, we maintain the assumption of Eq. (11) for the test data. This helps to compare the two settings with the same test samples. In fact, in practice, the data is most valuable when the class labels match for both image and text. Our dev and test sets for this setting are similar to the previous setting. However, the training set contains a larger number of samples where their image-text pairs are not necessarily labeled as the same class.

Setting C: This setting is closest to the real-world scenario where we analyze the new event of a crisis with a model trained on previous crisis events. First, we require the training and test sets to be from crisis events of a different nature (i.e., wildfire vs. flood). Second, we maintain the temporal component and only train on events that have happened before the tweets of the testing set. Since collecting annotated data on an urgent ongoing-event is not possible, and also because an event of crisis may do not have a similar annotated event in the past, these two restrictions often simulate a real-world scenario. For the experiments of this setting, there is no dev set. Instead, we use a random portion of the training data to tune the hyper-parameters.

Table 2. Number of samples in different splits of our settings.

Setting	# of Training samples	# of Dev samples	# of Test samples
Setting A			
Task1:	7876	553	2821
Task2:	1352	540	1467
Task3:	2590	340	358
Setting B			
Task1:	12680	553	2821
Task2:	5433	540	1467
Setting C			
Experiment 1:	174	-	217
Experiment 2:	4037	-	217
Experiment 3:	4761	-	217

We test on the tweets that are related to the California Wildfire (Oct 10 - 27, 2017), and train on the following three sets:

1. Sri Lanka Floods tweets (*May 31- Jul. 3, 2017*)
2. Sri Lanka Floods, and Hurricane Harvey and Hurricane Irma tweets (*May 31- Sept. 21 , 2017*)
3. Sri Lanka Floods, Hurricanes Harvey and Irma and Mexico Earthquakes (*May 31 - Oct. 5, 2017*).

4.3. Baselines

We compare our method against several state-of-the-art methods for text and/or image classification. There are a few categories of baseline methods we compare against. In the first category, we compare to DenseNet [30] and BERT [17], which are the state-of-the-art unimodal classification networks for images and texts respectively. We use Wikipedia pre-trained Bert and pre-train DenseNet on Imagenet [16]. The second category of baseline methods include several recently proposed multimodal fusion methods for classification:

- Compact Bilinear Pooling [22]: multimodal compact bilinear pooling is a fusion technique first used in visual question answering task but can be easily modified to perform standard classification task.
- Compact Bilinear Gated Pooling [38]: this fusion method is an adaption based on the compact bilinear pooling method for classification task by adding an extra gate on top.
- MMBT [37]: recently proposed supervised multimodal bitransformers model for classifying images and text, similar to VilBert [43] in some sense. We exclude VilBert because it does not work for image and text classification problem.

To further show the effectiveness of our model, we compare against score level *Score Fusion* and late feature fusion *Feature Fusion* of DenseNet and Bert networks. Score level fusion is one of the most common fusion techniques. It averages the predictions of separate networks trained on the different modalities. Feature Fusion is one of the most effective methods for integrating two heterogeneous modalities [50]. It concatenates deep layers

from modality networks to predict a shared output. We also provide three variations of our attention modules and report their performance: the first variant is to remove SSE; the second variant is to remove bottom SSE and atop self-attention in Figure 3; the third variant is to change our novel cross-attention to self-attention.

4.4. Evaluation Metrics

We evaluate the models in this paper using classification accuracy[4], Marco F1-score and weighted F1-score. Note that while in the event of a crisis, the number of samples from different categories often significantly varies, it is important to detect all of them. F1-score and weighted F1-score take both false positives and false negatives into account, and therefore, along with accuracy as an intuitive measure, are proper evaluation metrics for our datasets.

4.5. Training Details

We use pre-trained DenseNet and BERT as our image and text backbone networks. The details of their implementations can be found in [30] and [17], respectively. We do not freeze the pre-trained weights and train all the layers for both the backbone networks.

We use the standard SGD optimizer. We start with the base learning rate of 2×10^{-3} with a $10\times$ reduction when the dev loss is saturated. We use a batch size of 32. The models were implemented in Keras and Tensorflow-1.4 [1]. In all the applicable experiments, we select hyper-parameters with cross-validation on the dev set. For the experiments of the Setting 3 that we do not have an evaluation set, we tune hyper-parameters on 15% of the training samples. We select ρ and p_0 respectively in the range of $\rho \in [10, 20000]$ and $p_0 \in [0, 1]$.

We employ the following data augmentations on the images during the training stage. Images are resized such that the smallest side is 228 pixels, and then randomly cropped with a 224×224 patch. In addition, we produce more images by randomly flipping the resulting image horizontally.

For tweet normalization, we remove double spaces and lower case all characters. In addition, we replace any link in the tweet with the sentimental word "`link`".

5. Experimental Results

5.1. Setting A

As shown in Tables 3, 4, 5, our proposed framework SSE-Cross-BERT-DenseNet easily outperforms the standalone DenseNet and BERT models. We also compare our method against simple score fusion and feature fusion:

[4]In the settings that our experiments are defined classification accuracy is equivalent to Micro F1-score.

Table 3. Setting A: Informativeness Evaluation

Model	Dev Set			Test Set		
	Acc	Macro F1	Weighted F1	Acc	Macro F1	Weighted F1
DenseNet [30]	81.56	79.24	81.51	81.57	79.12	81.22
BERT [17]	83.91	82.19	84.46	84.90	81.19	83.30
Compact Bilinear Pooling[22]	87.88	86.15	87.86	88.12	86.18	87.61
Compact Bilinear Gated Pooling [38]	88.07	86.84	88.02	88.76	87.50	88.80
MMBT [37]	83.54	-	-	82.48	-	-
Score Fusion	87.88	86.55	87.80	88.16	83.46	85.26
Feature Fusion	86.62	86.33	87.67	87.56	85.20	86.55
Attention Variant 1	87.16	**87.98**	**89.28**	89.29	85.68	87.04
Attention Variant 2	87.52	86.89	88.32	88.34	86.12	87.42
Attention Variant 3	87.52	86.79	88.20	88.20	86.22	87.47
SSE-Cross-BERT-DenseNet	**89.51**	87.28	88.75	**89.33**	**88.09**	**89.35**

Table 4. Setting A: Humanitarian Categorization Task Evaluation

Model	Dev Set			Test Set		
	Acc	Macro F1	Weighted F1	Acc	Macro F1	Weighted F1
DenseNet [30]	87.22	55.85	82.67	83.44	60.45	86.96
BERT [17]	87.78	66.23	85.81	86.09	66.83	87.83
Compact Bilinear Pooling[22]	90.37	64.83	88.82	89.30	67.18	90.33
Compact Bilinear Gated Pooling [38]	89.44	60.12	85.32	85.34	65.95	89.42
MMBT [37]	89.25	-	-	85.82	-	-
Score Fusion	90.19	53.00	85.22	86.98	54.01	88.96
Feature Fusion	**91.48**	64.80	89.40	89.17	67.28	91.40
Attention Variant 1	90.93	59.96	87.69	88.41	64.60	90.71
Attention Variant 2	**91.48**	63.48	89.22	89.23	67.63	91.56
Attention Variant 3	90.37	61.11	86.67	87.18	64.67	90.24
SSE-Cross-BERT-DenseNet	91.14	**70.16**	**90.88**	**91.14**	**68.41**	**91.82**

Table 5. Setting A: Damage Severity Task Evaluation

Model	Dev Set			Test Set		
	Acc	Macro F1	Weighted F1	Acc	Macro F1	Weighted F1
DenseNet [30]	67.65	41.74	61.98	62.85	52.34	66.10
BERT [17]	67.06	40.69	62.29	68.16	45.04	61.09
Compact Bilinear Pooling[22]	71.47	48.50	65.53	66.48	**61.03**	**70.58**
Compact Bilinear Gated Pooling [38]	69.12	45.64	65.08	68.72	51.46	65.34
MMBT [37]	67.94	-	-	65.36	-	-
Score Fusion	71.47	49.72	67.11	71.23	53.48	66.26
Feature Fusion	65.88	37.72	60.13	67.60	40.62	56.47
Attention Variant 1	72.35	46.12	67.61	71.51	55.41	69.71
Attention Variant 2	70.29	43.89	62.79	63.13	58.03	69.39
Attention Variant 3	71.47	47.62	66.56	68.99	57.42	69.16
SSE-Cross-BERT-DenseNet	**74.58**	**58.80**	**72.55**	**72.65**	59.76	70.41

it is clear that our method enjoys an edge in every metric we used (Accuracy and F1 scores) for the test dataset. Compared with baseline methods Compact Bilinear Pooling [22], Compact Bilinear Gated Pooling [38], and MMBT [37], our proposed cross-attention fusion method does enjoy an edge over previous known fusion methods, including the standard score fusion and feature fusion. This edge holds true across Setting A, B and C. But at the same time, it is hard to tell which component plays a larger role in the performance: SSE, cross-attention and self-attention as this varies from task to task. We show the ablation results in Table 8.

One important observation we find across the three tasks is that despite the fact that accuracy percentages are reasonably good for simple feature fusion method, the macro F1 scores improve much more once we add attention mechanisms. We think this may be due to the fact that our proposed fusion method benefits the minority classes more than the majority class.

5.2. Setting B

In this setting, we want to see whether our models can perform better if we can make use of more labelled data

Table 6. Setting B: Informativeness Task and Humanitarian Categorization Task Evaluations

Model	Informativeness Task			Humanitarian Categorization Task		
	Accuracy	Macro F1	Weighted F1	Accuracy	Macro F1	Weighted F1
DenseNet [30]	83.36	80.95	82.95	82.89	66.68	83.13
BERT [17]	86.26	84.44	86.01	87.73	83.72	87.57
Score Fusion	87.03	85.19	86.90	91.41	83.26	91.36
SSE-Cross-BERT-DenseNet	**90.05**	**88.88**	**89.9**	**93.46**	**84.16**	**93.35**
Best from Table 3, 4, 5	89.33	88.09	89.35	91.48	67.87	91.34

Table 7. Comparing our proposed method with baselines for Humanitarian Categorization Task in Setting 3. We fix the last occurred crisis namely 'California wildfires' as test data and vary the training data which is specified in the columns.

Model	Sri Lanka Floods			Sri Lanka Floods + Hurricanes Harvey & Irma			Sri Lanka Floods + Hurricanes Harvey & Irma + Mexico earthquake		
	Accuracy	Macro F1	Weighted F1	Accuracy	Macro F1	Weighted F1	Accuracy	Macro F1	Weighted F1
DenseNet [30]	55.71	35.77	56.85	70.32	52.23	68.55	70.32	44.80	68.79
BERT [17]	31.96	20.90	27.21	73.97	53.90	73.51	74.43	56.98	74.21
Score Fusion	56.62	36.77	57.96	81.74	56.54	81.03	81.28	55.90	80.54
SSE-Cross-BERT-DenseNet	**62.56**	**39.82**	**62.08**	**84.02**	**63.12**	**83.55**	**86.30**	**65.55**	**85.93**

Table 8. Ablation Study of our proposed method for Humanitarian Categorization Task in Setting A.

Model	Test Set		
	Accuracy	Macro F1	Weighted F1
SSE-Cross-BERT-DenseNet	**91.14**	**68.41**	**91.82**
− Self-Attention	89.23	56.50	87.70
− Cross-Attention	88.48	56.38	87.10
− Cross-Attention + Co-Attention	88.55	60.47	87.92
− Cross-Attention + Self-Attention	86.30	58.33	85.27
− Dropout	83.37	54.83	82.46
− SSE	88.41	64.60	90.71
− SSE + Shuffling Within Class	88.68	62.91	88.33
− SSE + Mix-up [68]	89.16	54.63	87.37

for un-matched images and texts. Note that this involves training on noisier data than the prior setting. In Table 6, our proposed framework SSE-Cross-BERT-DenseNet beats the best results from Setting A for both the Informativeness Task (89.9 to 89.35 Weighted F1) and the Humanitarian Categorization Task (93.35 to 91.34). The gap between our method versus standalone BERT and DenseNet also gets wider. We believe the gap grows because it is very hard for standalone models to deal with covariate-shifts happening between the training and test datasets. On the other hand, our model seems to handle the covariate-shift situation between training and test data well, as one can easily seen when comparing Table 6 against Table 3, 4, 5. Note that the dev and test sets are the same for setting A and setting B while only the training data differs.

5.3. Setting C - Temporal

In this setting, we are curious to see whether our model still performs better than baseline models in a more realistic setting, i.e. the train / dev / test sets are split in the order they occured in the real world. We find that our proposed model consistently performs better than standalone image and text models (see Table 7). Additionally, the more crisis data we train on, the better the performance gets. There is a huge jump for almost every method (both baseline ones and ours) when training set contains one more crisis. This emphasizes the importance of collecting and labelling more crisis data even if there is no guarantee that the crises we collected data from will be similar to a future one. In the experiments, training crises contain floods, hurricanes and earthquakes but the test crisis is fixed at wildfires.

5.4. Ablation Study

In our ablation study, we examine each component of the model in Figure 3: namely self-attention on concatenated embedding, cross-attention on fusing image feature map & sentence embedding, dropout and SSE regularization. First, we find self-attention plays an important role on the final performance, accuracy drops to 89.23 from 91.14 if self-attention is removed. Second, the choice of cross-attention over co-attention and self-attention is well justified: we see the accuracy performance drops to around 88 by replacing the cross-attention. Third, dropout regularization [57] plays an important role in regularizing the hidden units: if we removes dropout completely, the performance drops to 83.37 from 91.14, which is indeed a large drop. Fourthly, we justify the usage of SSE [64] over the choice of Mixup [68] or within-class shuffling data augmentation. SSE performs better than mixup in terms of accuracy 91.14% versus 89.16%, and even much better in terms of F1 scores, 68.41 versus 54.63 for macro F1 score and 91.82 versus 87.37 for weighted F1 score.

6. Conclusions and Future Work

In this paper, we presented a novel multimodal framework for fusing image and textual inputs. We introduced a new cross attention module that can filter not-informative or misleading information from modalities and only fuse the useful information. We also used Stochastic Shared Embeddings (SSE) to regularize the training process and deal

with limited training data. We evaluate this approach on three crisis tasks involving social media posts with images and text captions. We show that our approach not only outperforms image-only and text-only approaches which have been the mainstay in the field, but also other multimodal combination approaches.

For future work we plan to test how our approach generalizes to other multimodal problems such as sarcasm detection in social media posts [13, 54], as well as experiment with different image and text feature extractors. Given that the CrisisMMD corpus is the only dataset available for this task and it is limited in size, we also aim to construct a larger set, which is a major effort.

References

[1] Mart'ın Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283, 2016. 7

[2] Rediet Abebe, Shawndra Hill, Jennifer Wortman Vaughan, Peter M Small, and H Andrew Schwartz. Using search queries to understand health information needs in africa. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 3–14, 2019. 1, 2

[3] Kashif Ahmad, Michael Riegler, Konstantin Pogorelov, Nicola Conci, Pål Halvorsen, and Francesco De Natale. 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*, page 12. ACM, 2017. 2

[4] Firoj Alam, Ferda Ofli, and Muhammad Imran. Crisismmd: Multimodal twitter datasets from natural disasters. In *Twelfth International AAAI Conference on Web and Social Media*, 2018. 5

[5] Mohamed R Amer, Timothy Shields, Behjat Siddiquie, Amir Tamrakar, Ajay Divakaran, and Sek Chai. Deep multimodal fusion: A hybrid approach. *International Journal of Computer Vision*, 126, 2018. 3

[6] Anonymous. {ALBERT}: A lite {bert} for self-supervised learning of language representations. In *Submitted to International Conference on Learning Representations*, 2020. under review. 4

[7] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 3

[8] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017. 3

[9] Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442, 2016. 3

[10] Terra Blevins, Robert Kwiatkowski, Jamie MacBeth, Kathleen McKeown, Desmond Patton, and Owen Rambow. Automatically processing tweets from gang-involved youth: Towards detecting loss and aggression. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2196–2206, Osaka, Japan, Dec. 2016. The COLING 2016 Organizing Committee. 2

[11] Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 136–145. Association for Computational Linguistics, 2012. 3

[12] Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and Joao Sedoc. Modeling empathy and distress in reaction to news stories. *arXiv preprint arXiv:1808.10399*, 2018. 2

[13] Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards multimodal sarcasm detection (an _Obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629, Florence, Italy, July 2019. Association for Computational Linguistics. 9

[14] JunKai Chen, Zenghai Chen, Zheru Chi, and Hong Fu. Emotion recognition in the wild with feature fusion and multiple kernel learning. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 508–513. ACM, 2014. 3

[15] Shizhe Chen and Qin Jin. Multi-modal dimensional emotion recognition using recurrent neural networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 49–56. ACM, 2015. 3

[16] Jia Deng, R. Socher, Li Fei-Fei, Wei Dong, Kai Li, and Li-Jia Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, volume 00, pages 248–255, 06 2009. 6

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1, 3, 4, 6, 7, 8

[18] Johannes C. Eichstaedt, H Andrew Schwartz, Salvatore Giorgi, Margaret L Kern, Gregory Park, Maarten Sap, Darwin R Labarthe, Emily E Larson, Martin Seligman, Lyle H Ungar, et al. More evidence that twitter language predicts heart disease: A response and replication. 2018. 2

[19] Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A Asch, and H Andrew Schwartz. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208, 2018. 2

[20] Golnoosh Farnadi, Jie Tang, Martine De Cock, and Marie-Francine Moens. User profiling through deep multimodal fusion. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 171–179. ACM, 2018. 3

[21] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013. 3

[22] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016. 3, 6, 7

[23] Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states. *Proceedings of the National Academy of Sciences*, 114(50):13108–13113, 2017. 1, 2

[24] Peter Gehler and Sebastian Nowozin. On feature combination for multiclass object classification. In *2009 IEEE 12th International Conference on Computer Vision*, pages 221–228. IEEE, 2009. 3

[25] Sharath Chandra Guntuku, Daniel Preotiuc-Pietro, Johannes C Eichstaedt, and Lyle H Ungar. What twitter profile and posted images reveal about depression and anxiety. *arXiv preprint arXiv:1904.02670*, 2019. 2

[26] Mihai Gurban, Jean-Philippe Thiran, Thomas Drugman, and Thierry Dutoit. Dynamic modality weighting for multi-stream hmms inaudio-visual speech recognition. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 237–240. ACM, 2008. 3

[27] Charles Harding, Francesco Pompei, Dmitriy Burmistrov, H Gilbert Welch, Rediet Abebe, and Richard Wilson. Breast cancer screening, incidence, and mortality across us counties. *JAMA internal medicine*, 175(9):1483–1489, 2015. 1, 2

[28] Jack Hessel, Bo Pang, Zhenhai Zhu, and Radu Soricut. A case study on combining asr and visual features for generating instructional video captions. *arXiv preprint arXiv:1910.02930*, 2019. 3

[29] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 1

[30] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 3, 4, 6, 7, 8

[31] Ilija Ilievski and Jiashi Feng. Multimodal learning and reasoning for visual question answering. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 551–562. Curran Associates, Inc., 2017. 3

[32] Andrew Ilyas. Microfilters: Harnessing twitter for disaster management. In *IEEE Global Humanitarian Technology Conference (GHTC 2014)*, pages 417–424. IEEE, 2014. 2

[33] Natasha Jaques, Sara Taylor, Akane Sano, and Rosalind Picard. Multi-task, multi-kernel learning for estimating individual wellbeing. In *Proc. NIPS Workshop on Multimodal Machine Learning, Montreal, Quebec*, volume 898, 2015. 3

[34] Xinyang Jiang, Fei Wu, Yin Zhang, Siliang Tang, Weiming Lu, and Yueting Zhuang. The classification of multi-modal data with hidden conditional random field. *Pattern Recognition Letters*, 51:63–69, 2015. 3

[35] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 3

[36] Stephen Kelly, Xiubo Zhang, and Khurshid Ahmad. Mining multimodal information on social media for increased situational awareness. 2017. 2

[37] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*, 2019. 6, 7

[38] Douwe Kiela, Edouard Grave, Armand Joulin, and Tomas Mikolov. Efficient large-scale multi-modal classification. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 3, 6, 7

[39] Shamanth Kumar, Geoffrey Barbier, Mohammad Ali Abbasi, and Huan Liu. Tweettracker: An analysis tool for humanitarian and disaster relief. In *Fifth international AAAI conference on weblogs and social media*, 2011. 2

[40] Zhen-Zhong Lan, Lei Bao, Shoou-I Yu, Wei Liu, and Alexander G Hauptmann. Multimedia classification and event detection using double fusion. *Multimedia tools and applications*, 71(1):333–347, 2014. 2, 3

[41] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018. 3

[42] Xukun Li, Doina Caragea, Huaiyu Zhang, and Muhammad Imran. Localizing and quantifying damage in social media images. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 194–201. IEEE, 2018. 2

[43] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019. 3, 4, 6

[44] Sreenivasulu Madichetty and M Sridevi. Detecting informative tweets during disaster using deep neural networks. In *2019 11th International Conference on Communication Systems & Networks (COMSNETS)*, pages 709–713. IEEE, 2019. 2

[45] Navonil Majumder, Devamanyu Hazarika, Alexander Gelbukh, Erik Cambria, and Soujanya Poria. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowledge-Based Systems*, 161:124–133, 2018. 3

[46] Hussein Mouzannar, Yara Rizk, and Mariette Awad. Damage identification in social media posts using multimodal deep learning. In *ISCRAM*, 2018. 2

[47] Ganesh Nalluru, Rahul Pandey, and Hemant Purohit. Relevancy classification of multimodal social media streams for

emergency services. In *2019 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 121–125. IEEE, 2019. 2

[48] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011. 3

[49] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*, 2018. 3

[50] Dhanesh Ramachandram and Graham W Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6):96–108, 2017. 6

[51] Geovany A Ramirez, Tadas Baltrušaitis, and Louis-Philippe Morency. Modeling latent discriminative dynamic of multi-dimensional affective signals. In *International Conference on Affective Computing and Intelligent Interaction*, pages 396–406. Springer, 2011. 3

[52] Tim GJ Rudner, Marc Rußwurm, Jakub Fil, Ramona Pelich, Benjamin Bischke, and Veronika Kopacková. Rapid computer vision-aided disaster response via fusion of multiresolution, multisensor, and multitemporal satellite imagery. 2

[53] Naina Said, Kashif Ahmad, Michael Regular, Konstantin Pogorelov, Laiq Hassan, Nasir Ahmad, and Nicola Conci. Natural disasters detection in social media and satellite imagery: a survey. *arXiv preprint arXiv:1901.04277*, 2019. 2

[54] Rossano Schifanella, Paloma de Juan, Joel Tetreault, and LiangLiang Cao. Detecting sarcasm in multimodal social platforms. *Proceedings of the 2016 ACM on Multimedia Conference - MM '16*, 2016. 9

[55] Himanshu Shekhar and Shankar Setty. Disaster analysis through tweets. In *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1719–1723. IEEE, 2015. 2

[56] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, pages 935–943, 2013. 3

[57] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 5, 8

[58] Nitish Srivastava and Ruslan R Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230, 2012. 3

[59] Kevin Stowe, Michael J Paul, Martha Palmer, Leysia Palen, and Kenneth Anderson. Identifying and categorizing disaster-related tweets. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 1–6, 2016. 2

[60] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019. 3

[61] Hien To, Sumeet Agrawal, Seon Ho Kim, and Cyrus Shahabi. On identifying disaster-related tweets: Matching-based or learning-based? In *2017 IEEE Third International Conference on Multimedia Big Data (BigMM)*, pages 330–337. IEEE, 2017. 2

[62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3

[63] Martin Wöllmer, Angeliki Metallinou, Florian Eyben, Björn Schuller, and Shrikanth Narayanan. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling. In *Proc. INTERSPEECH 2010, Makuhari, Japan*, pages 2362–2365, 2010. 3

[64] Liwei Wu, Shuqing Li, Cho-Jui Hsieh, and James Sharpnack. Stochastic shared embeddings: Data-driven regularization of embedding layers. *arXiv preprint arXiv:1905.10630*, 2019. 2, 3, 5, 8

[65] Yi Wu, Edward Y Chang, Kevin Chen-Chuan Chang, and John R Smith. Optimal multimodal fusion for multimedia data analysis. 2004. 3

[66] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019. 3, 4

[67] Jie Yin, Sarvnaz Karimi, Andrew Lampert, Mark Cameron, Bella Robinson, and Robert Power. Using social media to enhance emergency situation awareness. In *Twenty-fourth international joint conference on artificial intelligence*, 2015. 2

[68] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 8