Exploring Multimodal Features and Fusion Strategies for Analyzing Disaster Tweets

Raj Ratn Pranesh[1], **Ambesh Shekhar**[1], **Anish Kumar**[1]
[1]Birla Institute of Technology, Mesra
Ranchi, India
{raj.ratn18, ambesh.sinha, anishkr10052}@gmail.com

Abstract

Social media platforms such as Twitter often provide first-hand news during the outbreak of a crisis. It is extremely essential to process these facts quickly to plan the response efforts for minimal loss. Processing this social media information poses multiple challenges such as parsing noisy messages containing both texts and images. Furthermore, these messages are diverse, from personal achievements and opinions to situational crises.

Therefore, in this paper, we present an analysis of various multimodal feature fusion techniques to analyze and classify the disaster tweets into multiple crisis events via transfer learning. In our study, we utilized three image models-VGG19(Simonyan and Zisserman 2014), ResNet-50(He et al. 2016) and AlexNet pre-trained on ImageNet(Deng et al. 2009) dataset and three fine-tuned language models-BERT(Devlin et al. 2018), ALBERT(Lan et al. 2019) and RoBERTa(Liu et al. 2019) to learn the visual and textual feature of the data and combine them to make predictions. We have presented a systematic analysis of multiple intra-modal as well as cross-modal fusion strategies and their effect over the performance of the multimodal disaster classification system. In our experiment, we used 8,242 disaster tweets each consisting of image and text data with five disaster event classes. The results show that the multimodal with transformer-attention mechanism and factorized bilinear pooling (FBP)(Zhang, Wang, and Du 2019) for intra-modal and cross-modal feature fusion achieved the best performance.

Introduction

Sudden breakout of crisis generates a situation that is full of questions, uncertainties, and the obligation to make quick decisions, often with minimal news. Research in recent years has uncovered the importance of social media communications in disaster situations and shown that information broadcast via social media can improve situational awareness during a crisis (Vieweg et al. 2010). Social media has proved to be an active communication channel, especially during crisis events such as natural disasters like earthquakes, floods, typhoons ((Hughes and Palen 2009), (Imran, Mitra, and Castillo 2016)) or other emergencies such as accidents. These events spur a sudden surge of attention and actions from both the general public and the media. The early detection and analysis of such events are critical for the relief team. A quick breakdown of the crisis enables them to gain situational awareness and tactical information to effectively estimate early damage and to launch relief efforts accordingly.

An automated system for crisis-related information retrieval requires the extraction of relevant tweets and classifying them into pieces of types of information: affected individuals, infrastructural damages, casualties, donations, caution, or advice. Firstly, because the messages generated during a disaster vary greatly in value, an automatic system needs to filter out messages that do not contribute to situational awareness. These include the personal ones, irrelevant to the disaster. As a result, we design a system for detecting informative messages. Once the system detects the relevant tweets, it must classify these tweets to decide the type of information to extract (e.g. donation offers, casualty reports).

Information on social media mainly consists of textual messages and images. Past research has mainly focused on using textual content to aid disaster response. However, recent studies have revealed that images shared on social media during a disaster event can also help the relief team in several ways. For example, Nguyen et al. incorporated images shared on Twitter to assess the severity of infrastructure damage (Nguyen et al. 2017) in their work. Similarly, Jing et al. investigated the usefulness of images and text for their study on flood and flood aid (Jing et al. 2016). Our work addresses this and takes into account the information available from the text as well as images.

The **motivation** of our work to leverage the multimodal aspect of tweets for disaster event classification lies as follows: (i) each modality of the tweet carries a separate aspect, (ii) tweets with different modalities (text and image) can be used as separate features to maximize the entropy, (iii) the need for a quick and accurate multimodal framework to analyze the tweets for disaster relief efforts; models focusing on just one, visual or text feature is not enough to throw light on the magnitude of the situation completely.

Challenges: One of the foremost challenges here is to ascertain the nature of information present in a given tweet (e.g. donation offers, casualty reports, etc.). Labeling them in real-time would be difficult for humans themselves, let alone machines. So it should be perceived that this task re-

quires continuous work and effort if it is to be solved comprehensively. From a technical perspective, we face another challenge when working with tweets i.e. using the texts and images simultaneously. We would require deep and complex neural network architectures to achieve accurate results here. It poses another challenge, which is, analyzing the tweets fast-enough for the relief teams to make an early judgment.

In this paper, we presented an analysis of various multimodal fusion strategies for intra-modal fusion and cross-modal fusion. We investigated relation-attention, self-attention, and transformer-attention for intra-modal fusion. For the cross-modal fusion, we explored two methods, namely factorized bilinear pooling (FBP) and feature concatenation. Along with this, we evaluated state-of-the-art models which were three pretrained image models (VGG19, ResNet-50(He et al. 2016) and AlexNet) and three pretrained language models (BERT, RoBERTa and ALBERT) for the disaster tweet analysis and classification task. We found that the ResNet-50 outperformed other image models and among the textual models RoBERTa achieved the best performance. We further utilized these two models for the evaluation of intra-modal and cross-modal fusion strategies.

Our contributions in this paper can be summarized as follows:

- We present a systematic comparative study of various state-of-the-art image and language models for the task of understanding and classifying multimodal disaster tweets.
- We design and evaluate three variety of attention mechanisms based feature fusion strategies for combining linguistic and visual attributes.
- We investigated Factorized Bilinear Pooling (FBP) method and apply it for the cross-modal feature fusion.

RELATED WORK

The Harvard Humanitarian Initiative(2011) is one of the earliest works which provides a background on the Disaster Research Center and explains the strategic importance of studying the social science aspects of the disaster. An article by Palen and Liu (Palen and Liu 2007) was one of the first to provide an early assessment regarding how rapid advancements in communication technology can support the participation of the public during crises. Research in recent years highlights the use of social media by the public, formal response agencies, and other stakeholders during emergencies. These users interact in complex ways to produce, distribute, and organize the content (Starbird and Stamberger 2010). This enables tasks such as communicating about hospital availability (Starbird, Muzny, and Palen 2012), coordinating medical responses (Sarcevic, Marsic, and Burd 2012), and communicating with the public during various crises (Chan et al. 2014), among many others. Further information and a deeper perspective on users of social media in a disaster can be found in (Hughes et al. 2014) and (Hughes and Palen 2012). Multiple works have already shown that information that contributes to situational awareness is reported via Twitter (and other social media platforms) during mass emergencies ((Vieweg et al. 2010), (Vieweg 2012), (Imran et al. 2014)). (Chackungal et al. 2011) presents an in-depth analysis of the response to the earthquake in Haiti in January 2010. With a stronger focus on social media, a recent survey by Hughes, Peterson, and Palen focuses on the use of social media data by emergency responders to gain real-time notification of any emergencies. Hughes et al. (Imran et al. 2014) describe the challenges they face, best practices regarding the adoption of social media by formal response organizations, and also touch on instances of integrated, end-to-end systems that are currently being built to meet these needs. In recent year various work have be done using multimodal CrisisMMD(Alam, Ofli, and Imran 2018) dataset. In the paper(Ofli, Alam, and Imran 2020), author performed a multimodal classification for identifying disaster class and to decide if the data is informative or non-informative. Authors in paper(Agarwal et al. 2020)(Kumar et al. 2020) also performed classification task to classify multimodal data into two classes- informative and non-informative. The one common limitation of the previous work on multimodal classification is that their proposed models utilizes very simple feature fusion techniques. Not much exploration has been done for the improvement of the extracted visual and textual features and the enhancement of their combined multimodal representation. Secondly, most of the previous works focused on classification involving two classes(informative and non-informative). Paper(Ofli, Alam, and Imran 2020) performed disaster event classification but the number of classes were less. Papers such as (Ofli, Alam, and Imran 2020) performed disaster event classification but the number of classes were less.

In this paper, we addressed these limitation and presented an analysis of various feature fusion strategies for classifying multimodal disaster data. We utilized a multimodal dataset consisting of 8242 image-text disaster tweet pairs and classifies them into five fine-grained classes of crisis events.

METHODOLOGY

In this section, we elaborate on our multi-modal transfer learning framework as shown in Figure1. Our multimodal architecture can be divided into three sub-modules. The first sub-module consists of the extraction of contextual text feature using a language model. The second sub-module involves the extraction of the visual features from the image. Finally, at last, we have a fusion sub-module where feature representation received from the first and second sub-modal are combined to obtain a feature vector.

Textual feature extractor

In this section, we elaborate on the text feature extraction. This module deals with the extraction of contextual data from the tweets. We employ three pretrained language models namely-, BERT-base(Devlin et al. 2018), RoBERTa-base(Liu et al. 2019) and ALBERT-base(Lan et al. 2019) to extract high quality text feature vector. For the processing of textual data i.e. 'tweets' in the disaster dataset, we firstly performed the fine-tuning of each language model using the disaster tweets. We used the Hugging Face trans-

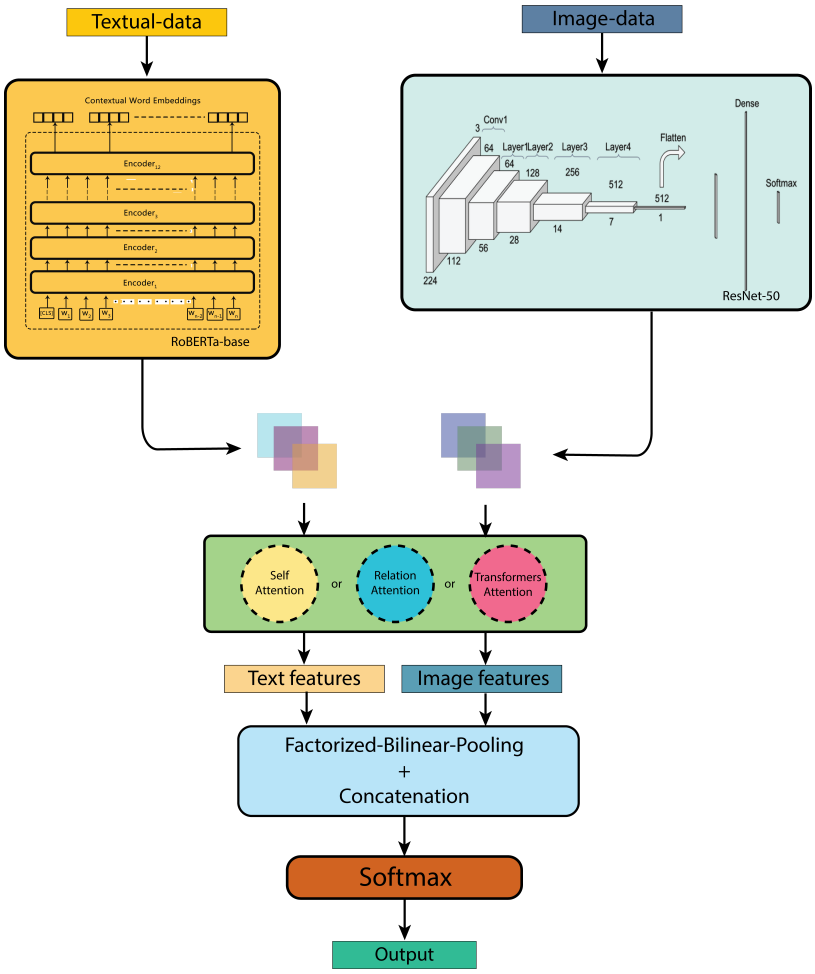

Figure 1: Feature fusion pipeline with textual sub-modal(RoBERTa on left) and visual sub-modal(ResNet-50 on right)

former model[1]. Each of the models consists of 12 encoding layers was used in this phase. For the finetuning of each model, we started by building a custom classification head on top of them. The classification head was consist of a dropout layer(p=0.05) followed by a linear layer(size = 768) with mish(Misra 2019) activation function followed by an another dropout layer and a final linear layer(size = 768). The averaged pool of sequential output from 12 encoding layers of each model was used as the custom classifier head's input. We used fast tokenizers[2] to efficiently tokenize and pad input text as well as prepare attention masks. Once the model gets fine-tuned, each of the language models was fed with a sequence of text inputs (tweets) which goes through all the stacked encoding layers, extracting essential features from the context.

As evident from the architecture Figure, for each language model, the encoders are stacked over each other having a sequence of tokens represented as w_i as input. The input flows up through the stacks and each layer applies self-attention and passes its results through a feed-forward network, and then hands it off to the next encoder. It outputs a $[CLS]$ or $< s >$ token of size equal to the hidden_size (768) for every input text sequence which is utilized as the textual feature representation. Finally, we collected three features vectors from each language model that holds meaningful information.

Visual Feature Extractor

In this section we describe the operation of the image feature extractor for which we used three image models namely-, VGG19(Simonyan and Zisserman 2014), ResNet-50 and AlexNet pretrained on Imagenet-21k(Deng et al. 2009). This module deals with the visual features of the image present in the disaster tweets. Images contain many thousands of pixel values in several colour channels, edges between two image regions, interest points and regions, ridges and their correlation and relationship characterizes the class and enables drawing a separation. These are some of the important features of an image which helps a deep learning model to learn better and hence, add to the classification accuracy of our model. A pretrained network is able to provide better re-

[1]Model is available at https://huggingface.co/models
[2]Available at https://github.com/huggingface/tokenizers

sults as compared to a convolution neural network developed from scratch.

In the architecture figure, to extract the essential visual features from the images, we supplied each of the three pre-trained image models with pre-processed image and generated visual representation was extracted from the final fully connected layer of each model. It outputs a vector of the dimension of 4096, 1000, 1000 VGG19, ResNet-50 and AlexNet respectively.

Multimodal fusion

In this section, we provided a detailed elaboration of both intra-modal and cross-modal fusion techniques utilized and investigated in this paper. For the fusion of intra-modal features of both textual and visual vector separately, we used three methods- self-attention, relation-attention and transformer-attention methods. For the cross-modal fusion of visual and textual features together, we used two methods- factorized bilinear pooling(FBP) and simple feature concatenation.

Intra-modal feature fusion We have developed functions using different attention-based methods, namely: self-attention, relation-attention and transformer-attention methods. These functions can convert a variable number of features into a fixed dimension feature. For n number of features, we denote the i_{th} feature as f_i where i $\in [1, n]. We applied fusion techniques as follows$:

- $Self-attention$: For each feature we apply a 1-dimensional fully connected layer $W^0_{d\times 1}$ and a sigmoid function σ, resulting to the weight a_i of the i_th feature f_i^T as follow:

$$\alpha_i = \sigma(f_i^T \cdot W^0_{d\times 1}) \tag{1}$$

  We combined these weights from self-attention for every feature into a global representation f_s as follows:

$$f_s = \frac{\sum_{i=1}^{n} \alpha_i f_i}{\sum_{i=1}^{n} \alpha_i} \tag{2}$$

- $Relation-attention$: The function derives the relationship between the features and generates relevant features. Since f_s holds global representation of these features, we use sample concatenation of each features and global representation to shape the global-local relation[f_i:f_s], next we apply the 1-dimensional fully connected layers $W^1_{d/times1}$ with the sigmoid function σ. For relation-attention weight β of i_th feature $[f_i : f_s]^T$ is computed as:

$$\beta_i = \sigma([f_i : f_s]^T \cdot W^1_{d/times1}) \tag{3}$$

  Using aggregated weights from self-attention function and relation-attention function, we combine all the features to get a new feature f_r:

$$f_r = \frac{\sum_{i=1}^{n} \alpha_i \beta_i [f_i : f_s]}{\sum_{i=1}^{n} \alpha_i \beta_i} \tag{4}$$

- $Transformer-attention$: After going through the works in (Zhang, Wang, and Du 2019) and (Yang et al. 2016), we compute the attention weight as

$$f'_i = \mathbf{W^{`}m}\times d^2 \cdot f_i + b\gamma_i = exp(\mathbf{u\hat{}t}_{d\times 1} \cdot tanh(f'_i)) \tag{5}$$

  To reshape the the dimension of feature f_i, we feed it through a $w \times d$ dimensional FC layer 5. The weight of i_{th} feature f_i is processed through tanh function which is then fed to the exp function along with dot product of $u^t_{d\times 1}$. We pass the output from the exp function to a 1-dimensional FC layer stated in 5. From the transformers attention we formulate all the features into a single feature f_i, as

$$f_s = \frac{\sum_{i=1}^{n} \gamma_i f_i}{\sum_{i=1}^{n} \gamma_i} \tag{6}$$

Cross-modal feature fusion In this section, we described two cross-modal fusion strategies used for combining the visual and textual feature vectors.

- $FactorizedBilinearPooling(FBP)$: The two feature vectors obtained via different modalities are fused together by applying Factorized Bilinear Pooling(FBP) function. The output from different modalities that is the text feature vector $\mathbf{a} \in R$ for textual data and image feature $\mathbf{v} \in R$, the cross-modal bilinear model is represented as

$$z_i = a^t W_i v \tag{7}$$

  where $\mathbf{z}_i \in R$ is the output from bilinear model, and $\mathbf{W} \in R$ is a projection matrix. The equation yields the output feature $z = [z_1, ..., z_0]$. Equation.7 and Equation.8 was formulated in paper(Zhang, Wang, and Du 2019).

$$z = [z_1, ..., z_0] = SumPooling(\tilde{U}^T a \circ \tilde{V}^T v, k) \tag{8}$$

  Figure(3) represents the equation.8, where by feeding features a and v to FC layer we got $\tilde{U}^T a$ and $\tilde{V}^T v$, respectively and the $SumPooling(x, k)$ applies sum pooling with non-overlapped windows to x. We add a dropout layer to prevent model from over-fitting. There may be variation in the output due to the element-wise multiplication in equation.8, therefore we applied the l2-normalization function to normalize the magnitude of output from dropout layer.

- $FeatureConcatenation$: The textual and visual feature vector obtained after the intra-modal feature fusion(i.e. $\mathbf{a} \in R$ and $\mathbf{v} \in R$ respectively) were then combined using simple concatenation technique to obtain the desired multimodal vector representation $\mathbf{m} \in R$.

$$\mathbf{m} = [\mathbf{a} \cdot \mathbf{v}] \tag{9}$$

(a) Hurricane Maria turns Dominica into 'giant debris field' https://t.co/rAISiAhMUy by #AJEnglish via @c0nvey https://t.co/I4zeuW4gkc

(b) A friend's text message saved Sarasota man from deadly California wildfire https://t.co/0TNMFgL885 https://t.co/CIzo44Npza

(c) Earthquake leaves hundreds dead, crews combing through rubble in #Mexico https://t.co/XPbAEIBcKw https://t.co/wGVxGD4xNd

Figure 2: Dataset Example:

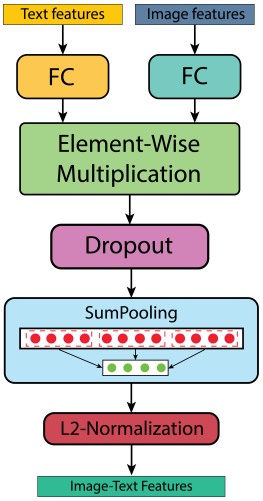

Figure 3: Cross-Modal Bilinear Model fusion

DATASET

In this paper, we have used CrisisMMD(Alam, Ofli, and Imran 2018) dataset for training and testing our model. The data distribution is shown in Table 1. This dataset consists of 8,242 tweets along with their associated images. Each text and image pair in the dataset have three annotations: (i) disaster event categories (eight classes), (ii) informative vs. not-informative, (iii) damage condition (three classes). In this paper, we have performed a classification using disaster event classes. Since the number of labels across different classes was uneven, we compressed the number of classes to five. The class 'Vehicle damage' was very small in number so we combined it with 'Infrastructure and utility damage' class. We also combined the 'Affected individuals'. 'Injured or dead people' and 'Missing or found people' to form one class- 'Affected individuals'. We labelled 'Not relevant or can't judge' as 'Non-Humanitarian' class. Other two classes were 'Rescue volunteering or donation effort' and 'Other relevant information'.

In the CrisisMMD dataset, tweet text and image in a pair were annotated separately, as a result, few pairs had a different label for text and it's associated image. We removed those pairs and performed the experiment only those data who have the same label for text and image. Finally, we split the data in 70%:15%:15% ratio which is equivalent to 5770:1236:1236 for training, development, and test sets, respectively.

Disaster Classes	**Total**	
	Text	Image
Not-humanitarian	4312	4312
other_relevant_information	1764	1764
rescue_volunteering_or_donation_effort	1195	1195
infrastructure_and_utility_damage	842	842
affected_individuals	129	129
Total	8242	8242

Table 1: Dataset Distribution

Text Preprocessing

The tweet data contained lots of noises and it was not suitable for the input in the model. We first cleaned and structured the tweet. For the preprocessing of tweets, we first created a list of Out-of-vocabulary (OOV) words which were replaced with meaningful complete words. Followed by removing URLs, blank rows, unwanted symbols, re-tweets and user-mentions. We used NLTK5[3], a Python module for text processing removed the English stopwords and performed lemmatization of tweets.

This textual data consists of a sequence of English words from which a maximum input sequence of 42 tokens are fed in our model. The max sequence length of text data was 52 but only 2% texts were greater than 42, so we decided to set max length limit to 42. If the sequence of tokens were greater than the limit, then the sequence was truncated otherwise padded from right respectively. The language models expect input data in a specific format. Therefore special tokens `<s>`, `</s>` for RoBERTa and `[CLS]`, `[SEP]` for BERT and ALBERT were added to mark beginning and separation or end of the sentences respectively. Now, for each model pre-trained tokenizer tokenizes the text and then replaces them with their respective input ID from tokenizer dictionary. A mask ID was also generated to distinguish between tokens and padded elements and a segment ID with positional embeddings was also needed to distinguish between different sentences and token position. Finally, we feed each language model with their respective input ID, mask ID and segment ID along with label ID.

[3]Available at https://www.nltk.org/

Image Preprocessing

For preparing the image input, we reshaped our image shape according to image model input_shape which was (224,224,3). We import images from a directory, followed by reshaping the size, these images were converted into arrays and every pixel value was then normalised before passing in the model.

EXPERIMENT

Experimental Setup of Baselines

For a comprehensive evaluation, we compare our model with the following baseline methods:

VGG16 + CNN(Ofli, Alam, and Imran 2020): This baseline model consists of VGG16 and CNN for image and text processing respectively. For visual features, we use the transfer learning approach. We use the weights of a VGG16 model pre-trained on ImageNet to initialize our model. The last layer (i.e., softmax layer) of the network is modified according to the particular classification task. For textual features, CNN with 5 hidden layers is used. The input is represented as a word-level matrix where each row represents a word in the tweet extracted using a pre-trained word2vec model. CNN layers are followed by max-pooling layer and finally one or more fully connected layers. The 1000-dimensional visual and textual vectors are then concatenated to form the 'Shared representation' and are passed through a hidden layer coupled with a softmax function to make final predictions. Adam optimizer is used for training the model. We use the early-stopping condition, and ReLU as an activation function on the weights to prevent overfitting.

Model	Precision	Recall	F1-score
AlexNet	74.42	56.74	64.38
VGG19	76.39	55.01	63.96
ResNet-50	79.23	60.11	68.35

Table 2: Performance of Image Unimodal

Exploring Visual feature

In the visual modal, we compared three image models, namely: AlexNet(Krizhevsky, Sutskever, and Hinton 2012),ResNet-50(He et al. 2016) and VGG19(Simonyan and Zisserman 2014); pretrained on large ImageNet(Deng et al. 2009) dataset. In the visual unimodal for each of the image model, the extracted feature vector was passed through two consecutive fully connected layers of dimension 512 and 256. The feature vector was then passed into a batch normalization layer and dropout layer(with dropout probability = 0.4), followed by a 5-dimensional dense layer with a softmax activation function in order to make the final class prediction of the disaster event. Relu activation function and L2 regularization of 0.01 was applied at each dense layer. All of the image models were trained on the training dataset(learning rate = 1e-4) using Adam(Kingma and Ba 2014) optimizer and with cross-entropy as the loss function. The model's hyperparameter fine-tuning was done on the validation set. We also conducted an evaluation of three models over the test dataset. As shown in the table 2, out of all three image models, ResNet-50 achieved the best F1 score of 68.35 as compared to ResNet-50(He et al. 2016) and AlexNet. This shows that the ResNet-50 was able to understand the image feature more clearly and generate better image representation. The reason behind this could be the residual module based ResNet-50's deeper architecture which lacks in VGG19 and AlexNet models.

Model	Precision	Recall	F1-score
ALBERT-base	77.34	66.02	71.23
BERT-base	79.34	67.47	72.92
RoBERTa-base	85.36	66.2	74.56

Table 3: Performance of Text Unimodal

Exploring Textual feature

Similar to visual modal, in the textual modal we have utilized the transfer-learning for learning the textual data representation. For the textual unimodal, we applied the bidirectional transformers with self-attention mechanism to extract resourceful features from the disaster tweets text. In our analysis, we used ALBERT-base(Lan et al. 2019), BERT-base(Devlin et al. 2018) and RoBERTa-base(Liu et al. 2019) pretrained language models. These models are mainly known for their pretrained weights over different domain data and for our task, we fine-tuned all of the models on the disaster dataset. As we discussed above, firstly, the input text sequence was structured, tokenized and pre-processed according to the language model's input format. From each of the language model, we extracted the $[CLS]$ (for BERT and ALBERT) or $< s >$ (for RoBERTa) which represent the entire input sentence and used as the aggregate sequence representation for classification tasks. Similar to the visual unimodal, the classification token was then passed through a series of the fully connected layer of size 512 and 256. Followed by a batch normalization layer, dropout layer(dropout probability = 0.4) and a 5-dimensional dense layer with a softmax activation function. All the dense layer in the model has a relu activation function and L2 regularization of 0.01. All of the models were trained with the learning rate of 1e-4, using Adam(Kingma and Ba 2014) as optimizer and cross-entropy as the loss function. On analysing the performance of all the three models on the test data, we observed 3 that the performance of RoBERTa-base unimodal was the best. BERT and ALBERT achieved the F1 score of 72.92 and 71.23 respectively.

Exploring Fusion Strategies

For exploring the intra-modal and cross-modal fusion techniques, as shown in the figure1, we used the best performing visual and textual model, ResNet-50 and RoBERTa.

Feature extraction: We extracted the feature maps from the preprocessed visual and textual data and utilized them

for the intra-modal fusion. For a given 3 dimension feature map, the size is represented as *HWC*, where *H* and *W* represented the height and width of the feature map respectively. The number of channel in the feature map was represented as C. For the intra-modal fusion process, we sliced the feature map into *n* vectors such that $n = H \ W$. Therefore, *n* number of C-dimensional vectors were obtained. For the image data, we extracted the feature map from the layer before the final average polling layer of the ResNet-50. For the RoBERTa model, instead of using classification token, we extracted the vector sequence consisting of each input token's vector representation. The size of each output token sequence was 768 x 42(max_length). This vector was split into 768 feature vector(42-dimensional) before intra-modal fusion.

Intra-modal Fusion: As we discussed above in the section *Multimodal Fusion*, we utilized 3 intra-modal attention fusion methods: relation-attention, self-attention, and transformer-attention. Both the visual and textual feature vector were subjected to each of the attention methods before performing the cross-modal fusion. The *n* split feature vectors from each of the visual and textual modalities, when passes through the attention layer, condenses to form respective unique representations which are then used for the cross-modal fusion.

Cross-modal Fusion: For the cross-modal fusion, we investigated 2 methods: factorized bilinear pooling (FBP) and feature fusion. The visual and textual feature vector generated after the intra-modal fusion is then subjected to cross-modal fusion to produce a combined multimodal representation. The multimodal vector is then passed through a classification layer of size 5 with a softmax activation function to make predictions. The model is trained on a batch size of 64 with cross-entropy loss function and Adam(Kingma and Ba 2014) optimizer for training the model. During the training of the model, we used an initial learning rate of 1e-5 two callback API- early-stopping condition and reduce learning rate on the plateau(reducing factor=0.5,patience=5).

TextualVisual	**Self attention**	**Relation attention**	**Transformers attention**
Self-attention	74.7%	75.4%	77.7%
Relation-attention	75.9%	78.8%	79.9%
Transformers-attention	78.0%	79.2%	80.1%

Table 4: Multimodal performance with feature fusion

TextualVisual	**Self attention**	**Relation attention**	**Transformers attention**
Self-attention	75.8%	78.1%	80.3%
Relation-attention	76.8%	79.3%	81.1%
Transformers-attention	79.1%	80.2%	85.5%

Table 5: Multimodal performance with FBP

RESULTS

In this section, we discuss and analyse the multimodal performance with various fusion technique. Table 4 and 5 shows the F1-score of various fusion methods we have experimented with. Table 4 shows the result of the multimodal framework after the simple feature fusion of the unimodal output, whereas table 5 shows the result of using factorized bilinear pooling for visual and textual feature fusion.

We clearly observed that by using the FBP(Zhang, Wang, and Du 2019) layer in the pipeline, the performance of multimodal was remarkably better(around **12%**) than the simple concatenation layer fusion. We also noticed that in either of the cross-modal fusion method, the transformer attention inta-modal fusion multimodal performed the best. In case of simple feature fusion, when same intra-modal fusion method was applied over the textual and visual data, the F1 score of self-attention, relation-attention and transfer-attention based multimodal were 74.7%, 78.8% and 80.1% respective. In case of the model with FBP, the scores were 75.8%, 79.3% and 85.5% respectively. We can also see that models having transformer-attention combined with relation-attention outperformed the model with transformer-attention and self-attention.

Coming to the multimodal baseline(Ofli, Alam, and Imran 2020), our model outperform it by **8.15%**. The reason behind the superior performance of our model lies behind the underlying language model and image models. Using transfer learning and attention-based fusion techniques, we were able to blend together with powerful language and image models and build a more robust multimodal.

CONCLUSION

In this paper, we presented an extensive analysis of multiple feature fusion strategies for developing a multi-modal framework for detecting and classifying tweets into various crisis events accurately based on the textual and visual features. In our study, we compared various transfer learning-based image and language models for the task and found that the ResNet and RoBERTa outperformed the other models. We also presented a comparative study of various fusion methods and through that, we can conclude that the selection of effective intra-modal and cross-modal method plays a very crucial role in developing a more accurate and efficient multimodal framework for classifying the events for faster relief efforts. We observed that the transformer-attention mechanism outperformed the other intra-modal fusion methods. We also showed that by using factorized bilinear pooling the multimodal feature representation can be improved. The experiments results show that one application of the multimodal framework can potentially be the identification and filtration of disaster-related information available on social media platforms but it is still far from perfect and there still exists room for improvement in the proposed design. Future work and possible experiments that can be done such as: (i) Experimenting with newer models for textual and visual feature extraction, (ii) Increasing the dataset size would definitely improve the performance, (iii) Providing additional features to the model would also enhance the performance.

References

Agarwal, M.; Leekha, M.; Sawhney, R.; and Shah, R. R. 2020. Crisis-DIAS: Towards Multimodal Damage Analysis-Deployment, Challenges and Assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 346–353.

Alam, F.; Ofli, F.; and Imran, M. 2018. Crisismmd: Multimodal twitter datasets from natural disasters. In *Twelfth International AAAI Conference on Web and Social Media*.

Chackungal, S.; Nickerson, J. W.; Knowlton, L. M.; Black, L.; Burkle Jr, F. M.; Casey, K.; Crandell, D.; Demey, D.; Di Giacomo, L.; Dohlman, L.; et al. 2011. Best practice guidelines on surgical response in disasters and humanitarian emergencies: report of the 2011 Humanitarian Action Summit Working Group on Surgical Issues within the Humanitarian Space. *Prehospital and Disaster Medicine* 26(6): 429.

Chan, J. C.; et al. 2014. 1 THE ROLE OF SOCIAL MEDIA IN CRISIS PREPAREDNESS, RESPONSE AND RECOVERY By .

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

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

Hughes, A. L.; and Palen, L. 2009. Twitter adoption and use in mass convergence and emergency events. *International journal of emergency management* 6(3-4): 248–260.

Hughes, A. L.; and Palen, L. 2012. The evolving role of the public information officer: An examination of social media in emergency management. *Journal of Homeland Security and Emergency Management* 9(1).

Hughes, A. L.; St. Denis, L. A.; Palen, L.; and Anderson, K. M. 2014. Online public communications by police & fire services during the 2012 Hurricane Sandy. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 1505–1514.

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

Imran, M.; Mitra, P.; and Castillo, C. 2016. Twitter as a lifeline: Human-annotated twitter corpora for NLP of crisis-related messages. *arXiv preprint arXiv:1605.05894* .

Jing, M.; Scotney, B. W.; Coleman, S. A.; McGinnity, M. T.; Zhang, X.; Kelly, S.; Ahmad, K.; Schlaf, A.; Gründer-Fahrer, S.; and Heyer, G. 2016. Integration of text and image analysis for flood event image recognition. In *2016 27th Irish Signals and Systems Conference (ISSC)*, 1–6. IEEE.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Kumar, A.; Singh, J. P.; Dwivedi, Y. K.; and Rana, N. P. 2020. A deep multi-modal neural network for informative Twitter content classification during emergencies. *Annals of Operations Research* 1–32.

Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; and Soricut, R. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* .

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* .

Misra, D. 2019. Mish: A self regularized non-monotonic neural activation function. *arXiv preprint arXiv:1908.08681* .

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*, 569–576.

Ofli, F.; Alam, F.; and Imran, M. 2020. Analysis of Social Media Data using Multimodal Deep Learning for Disaster Response. *arXiv preprint arXiv:2004.11838* .

Palen, L.; and Liu, S. B. 2007. Citizen communications in crisis: anticipating a future of ICT-supported public participation. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 727–736.

Sarcevic, A.; Marsic, I.; and Burd, R. S. 2012. Teamwork errors in trauma resuscitation. *ACM Transactions on Computer-Human Interaction (TOCHI)* 19(2): 1–30.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* .

Starbird, K.; Muzny, G.; and Palen, L. 2012. Learning from the crowd: Collaborative filtering techniques for identifying on-the-ground Twitterers during mass disruptions. In *ISCRAM*.

Starbird, K.; and Stamberger, J. 2010. Tweak the tweet: Leveraging microblogging proliferation with a prescriptive syntax to support citizen reporting. In *Proceedings of the 7th International ISCRAM Conference*, volume 1, 1–5. ISCRAM Seattle, WA.

Vieweg, S.; Hughes, A. L.; Starbird, K.; and Palen, L. 2010. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 1079–1088.

Vieweg, S. E. 2012. *Situational awareness in mass emergency: A behavioral and linguistic analysis of microblogged communications*. Ph.D. thesis, University of Colorado at Boulder.

Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; and Hovy, E. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 1480–1489.

Zhang, Y.; Wang, Z.-R.; and Du, J. 2019. Deep fusion: An attention guided factorized bilinear pooling for audio-video emotion recognition. In *2019 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.