A Multi-Layer Feature Low-Rank Fusion Algorithm Using Social Media for Disaster Information Detection

Bonan Li[1,a], Honglu Cheng[2,b], Jinyan Zhou[3,c], Rui Cao[4,d], Hong Zhang[5,e], Xingang Wang[6,f*]

[a]email: libonan@vip.163.com, [b]email: hulu6160@sina.com, [c]email: zhoujinyanwd@163.com
[d]email: caoruiqlu@163.com, [e]email: zhangqlu2022@163.com, [f*]email: xgwang@qlu.edu.cn.

[1]China Radio and Television Shandong Network Co, Ltd,China
[2,3,4,5,6]Key Laboratory of Computing Power Network and Information Security,Ministry of Education,Shandong Computer Science Center, Qilu University of Technology (Shandong Academy of Sciences) Jinan, China
[2,3,4,5,6]Shandong Engineering Research Center of Big Data Applied Technology, Faculty of Computer Science and Technology,Qilu University of Technology (Shandong Academy of Sciences) Jinan, China
[2,3,4,5,6]Shandong Provincial Key Laboratory of Computer Networks, Shandong Fundamental Research Centerfor Computer Science Jinan, China

Abstract. When disasters occur, people use social media to post in real time, which includes rich text and visual images. Relevant authorities can use this information to make emergency decisions and public opinion analysis quickly. However, the high complexity of multimodal deep learning models cannot meet the high timeliness requirements of disaster analysis. In addition, social media multimodal datasets for disaster detection are often scarce, and simple features are not sufficient to adequately provide usable information for analysis models. To mitigate these concerns, this paper introduces a model for low-level fusion called the Multilayer Feature Low-Level Fusion Model (MLLMF). The model uses transfer learning pre-trained models to extract text features from different hidden layers instead of traditional single-layer text features, and introduces gate attention units (GAUs) to enhance each modal feature to fully extract the intrinsic information of each single modality so that the information of small sample data can be utilized to the maximum. Moreover, to tackle the challenge posed by the high complexity of multimodal models, this paper uses low-rank tensor for multimodal fusion of text and images. With low model complexity, not only the unique information of a single modality is retained, but also the correlation between different modal elements is exploited to achieve a balance between low model complexity and high accuracy. Experiments show that the method not only improves processing efficiency but also achieves higher accuracy compared to unimodal and strong multimodal baseline methods, making it more suitable for crisis-related tasks.

Keywords: Multi-modal; social media; low-rank fusion; multi-layer feature extraction

1 Introduction

Natural disasters that occur annually worldwide often have the characteristics of suddenness, complexity, and dynamism[1]. The occurrence of natural disasters can cause significant human casualties and serious economic losses[2], so timely and effective disaster emergency management is of great significance in reducing disaster losses[3].

PMBDA 2023, December 15-17, Nanjing, People's Republic of China

DOI 10.4108/eai.15-12-2023.2345422

Social media, as a type of spatiotemporal big data, with its real-time and location service characteristics, has become one of the main research objects of disaster management[4,5]. Leveraging social media data analysis technology in disaster emergency plans and crisis management can significantly enhance the efficiency of emergency response coordination[6].

Previously, the disaster response research community employed traditional machine learning techniques to automate response activities [7], but reliance on manual features delayed model application, making the research time-consuming [8]. Recently, deep learning, which relies less on manual features, has gained popularity for acquiring advanced representations directly from input data, prompting exploration of its potential in disaster recovery activities.

Currently, there is little research on the automatic detection of crisis events using a combination of visual and textual modality information. On the one hand, only certain few-sample datasets (CrisisMMD)[9] are available due to the expensive acquisition of multimodal labeled datasets, and on the other hand, although the multimodal feature fusion approach can increase the model detection accuracy, the different modal data structures are different and may suppress each other or even be inferior to unimodal deep learning models without considering the asymmetry of their data structures. In addition, many powerful pre-trained models and high-complexity fusion algorithms lead to models with surprisingly large sizes, which further leads to a significant rise in runtime. Therefore, The objective of this study is to take into account three aspects related to the size of the model, detection accuracy, and the utilization of limited sample data to obtain an efficient multi-modal disaster monitoring model that balances accuracy and speed.

This article proposes a framework that combines image and text information to detect crisis events. In particular, this study suggests an automated labeling method for image-text pairs, focusing on the following criteria/tasks: 1) Informativeness, determining whether social media posts serve as informative tweets that contribute to facilitating humanitarian aid during disasters; 2) Event Classification, identifying the type of emergency situation conveyed by the post. This framework includes several steps, in which a fine-tuned image pre-training model and a text pre-training model are used as single-modal feature extractors given an image-text pair. Finally an improved low-rank fusion framework is proposed. Ultimately, the model obtains faster operational efficiency with the same accuracy.

In brief, this paper introduces a novel multi-modal framework designed for the classification of multi-modal data within the crisis domain. The contributions of this study include:

1. Proposes a multilevel feature representation of text using a pre-trained model, which can achieve higher accuracy than using the final output layer of the model with less sample fine-tuning.

2. A lightweight low-rank fusion method is presented in this work, considering both the unimodal specificity information and the symmetric intrinsic information association within the multi-modal structure. The optimization of this method aims to achieve a balance between accuracy and speed.

2 Related Work

Social media has been widely recognized as one of the most relevant and diverse resources, with numerous applications in identifying emergency health hazards[10], screening and detecting natural disasters[11–13], or depicting instances of violence and aggression in social media[14]. Previous studies of social media for detecting crisis events have focused on mostly text-based messages. For example, Shekhar et al.[15] formulated a crisis analysis system aimed at evaluating the extent of damage to assets and the degree of suffering experienced by victims. Sitaula et al.[16] proposed an end-to-end model integrating three feature extraction methods to analyse people's emotions during the COVID-19 epidemic.

The use of images for disaster detection has been an active frontier, both for user contributed content and satellite imagery (for a survey, see Said et al.[11]). For example, Li et al.[17] applied visualization methods and convolutional neural networks to identify and assess damage in images related to disasters. Nalluru et al.[18] integrated textual semantics and image features to classify social media tweets.

A deep multi-modal learning framework can be employed to consolidate complementary information derived from various modalities of the same phenomenon. Nevertheless, there is a relatively limited availability of multi-modal learning frameworks in the crisis domain. Currently, there are relatively few crisis datasets available, and one of the only multimodal crisis datasets, CrisisMMD[9], also exhibits a small sample size. This problem leads to a simple early[19] or late fusion[20] approach that may inhibit inter-modal interactions and thus lose the unique context and time dependence of each modality, making it difficult to train multi-modal models with strong and efficient generalization capabilities. For example, [22] introduced a tensor fusion network that calculates the outer product between single-peaked representations of three distinct modalities to derive a tensor representation. These methods use tensor representation to model multi-modal interactions and have achieved significant results. However, the computational complexity of this approach grows exponentially. [23] suggested a generalized low-rank multimodal approach based on the low-rank tensor approximation method, which incorporates significantly fewer model parameters and computational complexity compared to the tensor fusion method.

Based on the current crisis dataset and research status, the multilevel feature low-rank fusion framework introduced in this research optimizes the problems of poor generalization ability and high computational complexity of efficient models for small datasets. The specifics of the model are elaborated in the subsequent section.

3 Methodology

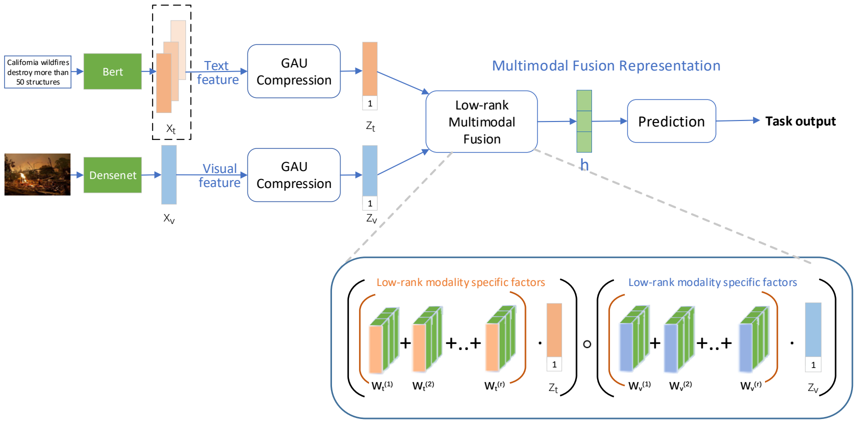

Figure 1: Overall framework of MLLMF.

The architecture proposed in this study is tailored for classification problems that involve inputting image-text pairs, such as user-generated tweets on social media, as illustrated in Figure 1, Here, the DenseNet and BERT graphs are sourced from [24] and [25], respectively. The method comprises four components: the initial two parts extract feature mappings from images and multilevel embedding representations from text, respectively; the third part reinforces text and image features through GAU[26] gated attention units, respectively; the fourth part includes a low-rank fusion of image features and text embedding. Each module will be described in later subsections.

3.1 Image Model for Feature Map Extraction

For images, they are data augmented and used as input, and feature mapping is extracted from the images using convolutional neural networks (CNNs). DenseNet[24] was chosen for this model, as it diminishes the size of the modules and amplifies the connections between layers. This choice is made to tackle parameter redundancy and enhance accuracy.

Thus, for each image v_i :

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

Here, v_i represents the *i*-th input image, $f_i \in \mathbb{R}^{D_f}$ is the vectorized form of the depth feature map in DenseNet, and the dimension $D_f = W \times H \times C$, where W, H, and C represent the height, width, and number of channels of the feature map, respectively.

3.2 Embedding Extracted Text Model

Considering the few-sample feature of the crisis dataset, the full utilization of features can make the model have a stronger discriminative basis to improve the generalization ability and robustness of the model.. Several studies[16] have shown that multiple embeddings can improve the characterization of text data and achieve higher accuracy. Inspired by this idea, the multiple hidden layers of BERT model becomes a breakthrough point. The BERT model consists of a 12-layer coding network with a hidden state (hidden) size of 768 for each layer. Where different hidden layers encode different linguistic hierarchical information[27]: the bottom layer network produces surface information features, the middle layer network generates syntactic information features, and the higher layer network yields semantic information features. For this purpose, the classification effects of different levels of features and the joint discriminations of multi-level features are experimented separately.

This paper uses monolingual models (BERT-BASE) as the core model for extracting embeddings from text. And the BERT model is used to pre-train the crisis-related tweet data from Wiki and Books. For each text input t_i .

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

Here, t_i represents the i-th sequence of word-piece tokens and $e_i \in \mathbb{R}^{756}$ is the sentence embedding. Following the approach outlined in the BERT paper [8], this study utilizes the embedding associated with[CLS] represent the entire sentence.

Detailed descriptions of the additional improvements made to the features extracted from DenseNet and BERT are provided in the subsequent section.

3.3 Gated Attention Unit (GAU) Intra-modal Feature Enhancement

Different modalities have different feature forms, and the feature fusion process is prone to the loss of single-peak information. Transformer[28] is a powerful sequence model that has achieved great success in feature extraction and enhancement - whether text, image or sound. However, its time and memory requirements grow second order with sequence length. For this reason, the Gated Attention Unit (GAU) introduces a gating mechanism to reduce the burden of self-attentiveness in the transformer. It combines the Gated Linear Unit (GLU) and the attention mechanism. The relevant layers are first introduced:

Vanilla MLP. Let $X \in \mathbb{R}^{T \times d}$ be the representation on the T marker. The output of the transformer MLP can be expressed as $O = \phi(XW_u)W_o$, where $W_u \in \mathbb{R}^{d \times e}$ and $W_u \in \mathbb{R}^{d \times e}$.

In this context, d represents the model size, e denotes the extended intermediate size, and ϕ is the element activation function.

Gated Linear Unit (GLU) is an enhanced Multi-Layer Perceptron (MLP) incorporating gating mechanisms [29]. The effectiveness of GLU has been demonstrated in various scenarios, and it is widely employed in state-of-the-art Transformer language models.

$$U = \varphi_u(XW_u), \quad V = \varphi_v(XW_v) \in \mathbb{R}^{T \times e}, \tag{3}$$

$$O = (U \odot V)W_o \in \mathbb{R}^{T \times d}, \tag{4}$$

In GLU, the original input vector X is assigned different weight matrices Wu and Wv as shown in Equation 3. In Equation 4, $\odot$ denotes element multiplication.

The Gated Attention Unit (GAU) is conceptualized to merge attention and GLU into a unified layer, aiming to maximize the sharing of computations between them. This approach not only enhances parameter and computational efficiency but also inherently incorporates a potent attention-gating mechanism. Specifically, GAU generalizes Equation 4 in GLU as follows:

$$O = (U \odot V')W_o \text{,} \tag{5}$$

where $V' = AV$ and $A \in \mathbb{R}^{T \times T}$ contains token-token attention weights. In contrast to GLU, which consistently employs v_i to select the pass u_i, GAU substitutes v_i with a potentially linked attentional representation $v_i' = \sum_j a_{ij} v_j$, this representation is retrieved from all available tokens to obtain a more relevant and contextually informed representation.

The presence of gating allows the use of simpler/weaker attention mechanisms than MHSA without loss of quality.

$$Z = \varphi_z(XW_z) \in \mathbb{R}^{T \times s}, \tag{6}$$

$$A = \mathrm{ReLU}^2(Q(Z)K(Z)^T + b) \in \mathbb{R}^{T \times T}, \tag{7}$$

In this expression, Z represents the shared representation, Q and K are two inexpensive transformations that apply to each dim scalar and the offset of Z, and b is the relative position deviation. The attention weights A are obtained by activating the transformed matrix.

GAU not only has the parallel computational features of GLU to speed up the model efficiency but also preserves the interaction between different features within the modality, capturing the importance of different locations in the text and features of different regions in the image. Each GAU layer has fewer parameters compared to the Transformer layer. More importantly, its quality is less dependent on attentional accuracy and is comparable to Transformer's performance.

3.4 Low-rank Fusion multi-modal Features

LMF is a tensor fusion method that models single-peak and two-peak interactions without using expensive Cartesian products[22]. Instead, the method uses single-peak features and weights to directly approximate the complete multi-tensor outer product operation. Tensor fusion takes into account the problem of structural asymmetry and semantic irrelevance between heterogeneous data. The elements of different modal feature representations are multiplied two by two to eliminate the unbalanced distribution of different modalities and obtain symmetric fusion features. This low-rank matrix decomposition operation can easily be extended to problems with very large interaction spaces (feature spaces or number of modes). In this paper, we use the approach described in[23]. Unlike its work, the model uses 3-layer GAU to augment individual mode-specific information and use it for mode-specific fusion, achieving model efficiency gains at a much smaller scale. Figure 1 depicts the improved LMF method, similar to the illustration in[23].

$$Z = \begin{bmatrix} z_v \\ 1 \end{bmatrix} \otimes \begin{bmatrix} z_a \\ 1 \end{bmatrix} = \begin{bmatrix} z_v & z_v \otimes z_l \\ 1 & z_l \end{bmatrix}, \tag{8}$$

$$\begin{aligned} h &= \left(\sum_{i=1}^{r} \otimes_{m=1}^{M} w_m^{(i)}\right) \cdot Z \\ &= \sum_{i=1}^{r}(\otimes_{m=1}^{M} w_m^{(i)} \cdot Z) \\ &= \sum_{i=1}^{r}(\otimes_{m=1}^{M} w_m^{(i)} \cdot \otimes_{m=1}^{M} z_m) \\ &= \Lambda_{m=1}^{M}[\sum_{i=1}^{r} w_m^{(i)} \cdot z_m] \end{aligned}, \tag{9}$$

Equation 8 shows how to add 1s to the single-peak tensor sequence before taking the outer product to equivalently capture the tensor representation of single-peak and multi-peak interactions, and the compressed representation (h) in Equation 9 is calculated using a low-rank mode-specific factor Wm and a batch Zm matrix multiplication of the additional modal representations. The product of all low-rank products is further multiplied to obtain the fusion vector.

4 Experimental Setup

4.1 Dateset

Crisis datasets are notably scarce, with the CrisisMMD [9] standing as the sole multi-modal crisis dataset currently available. This dataset comprises annotated image-text pairs extracted from tweets, where images and texts are labeled independently. The dataset was collected using event-specific keywords and hashtags during seven natural disasters in 2017, namely, hurricane Irma, hurricane Harvey, hurricane Maria, Mexico earthquake, California wildfires, Iraq-Iran earthquake, and Sri Lanka floods. The annotations in the corpus encompass two types:

Task 1: Informational or Non-informational: Given a tweet text or image, whether it contains information useful for humanitarian assistance.

Task 2: When presented with an image or tweet, the task is to classify it into one of the following five categories:

•Damage to infrastructure and utilities

•Vehicle damage

•Rescue, volunteer or donation efforts

•Affected individuals (injured, killed, missing, found, etc.)

•Other relevant information

Task 3: Damage severity: The evaluation involves assessing and categorizing the severity of damage depicted in the tweeted images as severe, minor, or minor/none.

Note that the annotation of the last task is only on the images. In this paper, we perform task 1 and task 2 (text-only, image-only and combined) on this dataset.

4.2 Settings

The tweet images and texts in the CrisisMMD dataset are independently tagged. Consequently, there are cases where images and texts in the same pair may correspond to different tags for Task 1 or Task 2. To ensure a comprehensive evaluation and offer practical insights, three evaluation settings were conducted, following the approach outlined in [30]:

Setting A: Only image-text pairs with consistent image and text labels are selected. The data from all seven crisis events were mixed, and the dataset was split into a training set, a validation set, and a test set.

Setting B: Image-text pairs with different labels are included in the training set, while the test set remains the same as in Setting A.

Setting C: This setting simulates a realistic crisis tweet classification task, where the model is trained only on events that occurred before the events in the test set. This approach closely resembles a practical scenario where a model trained on previously occurring crisis events is used to analyze a new crisis event.

Table 1 displays the number of samples per group for different settings and tasks.

Table 1: The quantity of samples in each of our settings' splits.

Setting	# of Training samples	# of Dev samples	# of Test samples
Setting A			
Task1:	7849	546	2798
Task2:	1346	534	1462
Setting B			
Task1:	12647	546	2898
Task2:	5427	534	1462
Setting C			
Experiment 1:	173	-	215
Experiment 2:	4031	-	215
Experiment 3:	4756	-	215

4.3 Baselines

This study contrasts this strategy with a number of cutting-edge techniques for text and/or picture categorization. Experiments compare it to the most widely used image and text single-peak classification networks, DenseNet and BERT, respectively, in the first category. In this research, we fine-tune the pre-trained DenseNet and BERT models from Wikipedia using the training data.

Several recently suggested multi-modal fusion approaches for classification are included in the second class of baseline methods:

- Compact Bilinear Pooling[21]: Initially developed for visual quizzing problems, the multi-modal compact bilinear pool is a fusion approach that may be readily adapted to carry out common classification tasks.

- Compact Bilinear Gated Pooling[31]: This fusion method represents a modification of the compact bilinear pooling approach, incorporating additional attention-gates into the compact bilinear pooling module.

- MMBT[32]: A supervised multi-modal bi-directional converter model for text and picture classification has been suggested lately.

- TFN[22]: A large multidimensional tensor fusion network based on sentiment analysis proposed for fusion across different modalities.

- SSE-Cross-BERT-DenseNet[30]: uses a multi-modal graph-based technique to create fresh matching pairs from various samples in order to deal with a limited amount of training data.

DenseNet and BERT networks for feature fusion, together with score level fusion and late feature fusion, comprise the third category. Of all the fusion procedures, score-level fusion is the most often used. It takes the average of separate networks' predictions that were trained on various modalities. One of the best techniques for combining two modalities is feature fusion[33]. In order to forecast common outputs, it links the deep layers of modal networks.

The MLLMF model proposed in this paper is compared with the baseline model described above.

4.4 Training Details

Text and image backbone networks, respectively, are pre-trained DenseNet and BERT, which are refined with text-only and picture-only training data. Implementation details are available in [24] and [6], correspondingly. All layers for the two backbone networks are trained by the framework, which does not freeze the pre-learned weights. For the text, the input layer is assumed to be h0 , the twelve encoder layers are $h_1, h_2, h_3, ..., h_{12}$. In this study, we evaluate the classification effectiveness by choosing different hidden layer features for crosstalk based on the informativeness task under setting A, as discussed in Section 4.2. The experiments confirm the effectiveness of our proposed combination of surface information features, syntactic information features, and semantic information features. The experimental results are presented in Table 2.

It is experimentally demonstrated that the best Accuracy values are obtained using the joint features of the bottom, middle and top layers concatenating. Therefore, in this paper, the output of BERT’s 1st, 5th, 9th and 12th hidden layers are concatenated as the embedded representation of the text.

The standard SGD optimizer is used in this paper. We start with a base learning rate of 2 × 10−3 and reduce it by a factor of 10 when the validation loss saturates. The batch size is 32.

Table 2: Task evaluation of features of different coding layers

Feature	Notation	F1 score
Embeddings	h0	74.62
Second to last hidden	h11	78.57
Last hidden	h12	80.67
Concat last four hidden	h9 to h12	80.83
Skip extraction four hidden	h1,h5,h9,h12	81.16
Concat all 12 hidden	h1 to h12	80.69

In all applicable experiments, We all use the validation set to select the best hyperparameters using the cross-validation method.

For experiments without validation sets at setting C in section 4.2, the experiments adjust the hyperparameters on 15% of the training samples.

The subsequent image data expansion is carried out in this paper during the training stage. After resizing the image to achieve a minimum edge of 228 pixels, 224 × 224 face slices are randomly used for cropping. Furthermore, by rotating the created pictures horizontally at random, this study produces additional images.

5 Experimental Results

5.1 Setting A: Leave Out Training Pairs with Contradictory Labels

In this instance, it is examined if annotated inconsistent text and photos are utilized as labeled data to see if the current model performs better. Keep in mind that compared to the prior configuration, this requires training with noisier data. Figure 2 shows that the MLLMF framework suggested in this study performs better than the best results of setting A in the humanitarian classification test (92.67 to 91.94) and the informative task (89.61 to 89.36 weighted F1). When compared to other multi-modal fusion frameworks, this technique still produces good results.

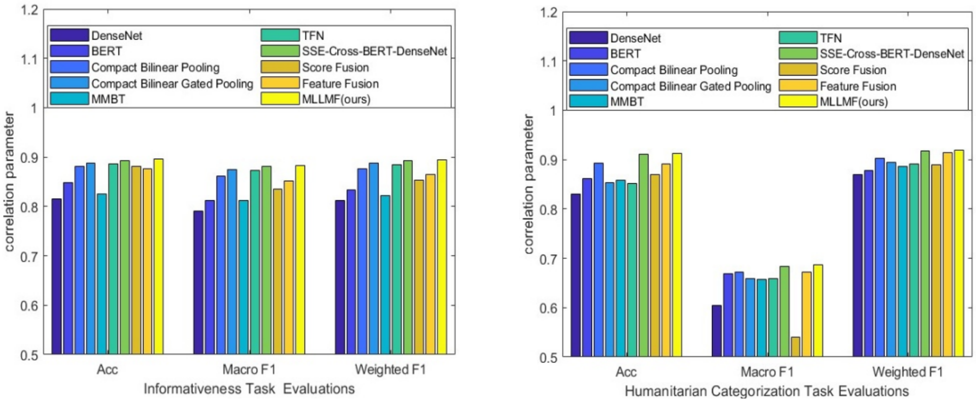

Figure 2: Setting A: Task Evaluations for Humanitarian Categorization and Informativeness.

The MLLMF model uses a lighter GAU module than the transformer to replace the LSTM module in the original LMF method, improving accuracy while reducing model size. The module used is faster, has a lower memory footprint, and is more effective than other models with comparable accuracy. Consistent speedups can be observed using this method of training with the same infrastructure and resources. The average time (in seconds) per batch measured at a fixed batch size compared to models such as the compact bilinear pool[21], compact bilinear gated pool[31], and TFN[22] with comparable accuracy under setup A is shown in Table 3, while Table 4 shows that the model uses a smaller number of trainable parameters.

An notable finding in both jobs is that, after the GAU module is applied, the macro F1 scores significantly increase, even while the accuracy percent ages are appropriate for basic feature fusion approaches.

Table 3: Average Time/Epoch (sec)

Model	Informativeness Task	Humanitarian Categorization Task
Compact Bilinear Pooling[21]	74.86	75.32
Compact Bilinear Gated Pooling[31]	80.64	81.57
TFN[22]	88.46	90.42
SSE-Cross-Bert-DenseNet[30]	64.35	68.74
MLLMF	47.36	47.84

Table 4: Number of Model Parameters

Model	Informativeness Task	Humanitarian Categorization Task
Compact Bilinear Pooling[21]	6487458	6756485
Compact Bilinear Gated Pooling[31]	6842547	6915486
TFN[22]	7438687	7635733
SSE-Cross-Bert-DenseNet[30]	5732945	5876542
MLLMF	3565485	3626573

5.2 Setting B: Incorporate Training Pairs with Varying Labels

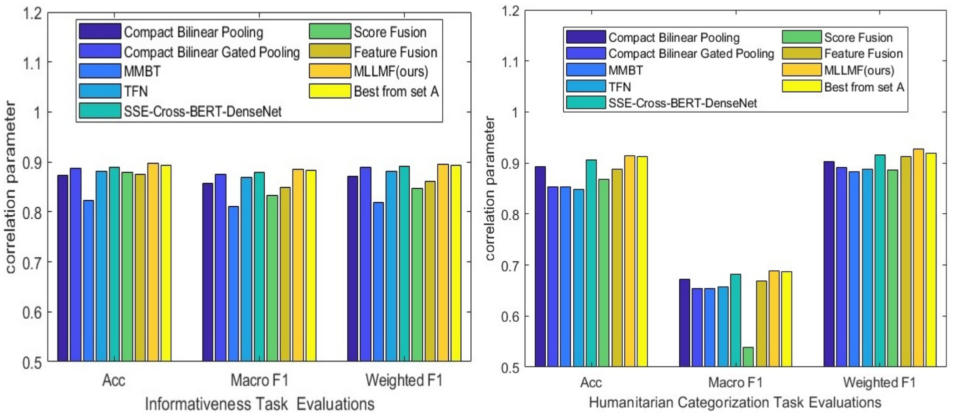

Figure 3: Setting B: Task Evaluations for Humanitarian Categorization and Informativeness.

In this case, it is investigated whether the present model performs better if labeled inconsistent images and text are used as labeled data. In Figure 3, the MLLMF framework proposed in this paper outperforms the best results of setting A for both the informative task and the humanitarian classification task. This method still achieves good results compared with other multi-modal fusion frameworks.

5.3 Setting C: Temporal

The purpose of this arrangement is to replicate a real-world situation, where the data used have the appearance of being solely historical (training/testing sets are arranged chronologically according to the real-world occurrence dates). Moreover, the data used for testing and training are not related to the same problem. As Figure 4 illustrates, the results show that the model suggested in this study regularly performs better than alternative multi-modal fusion frameworks.

Despite the fact that there is no assurance that the crises for which data are gathered would be comparable to other crises in the future, the findings highlight the need of gathering and classifying more crisis data. Floods, hurricanes, and earthquakes were among the training crises in the trial; however, the test crisis was limited to recognizing wildfires.

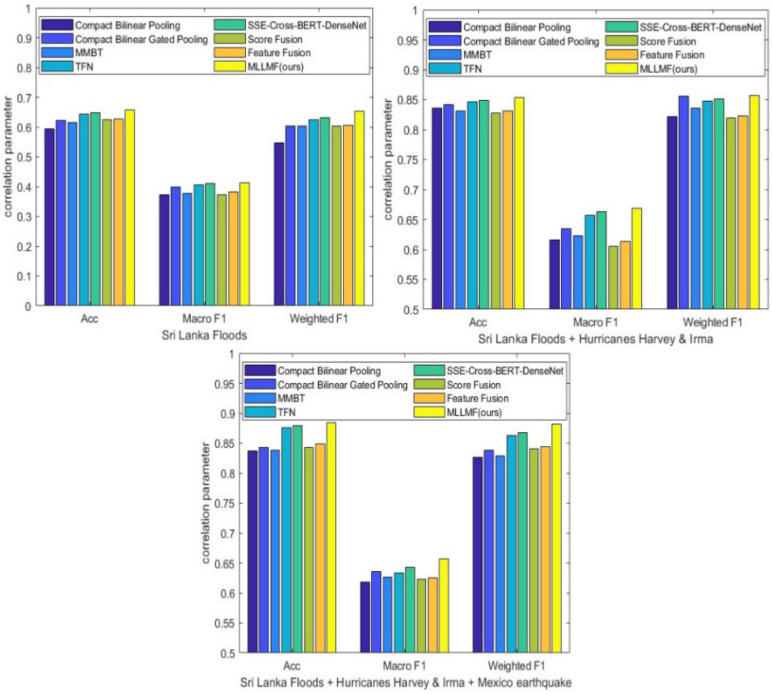

Figure 4: contrasting the Humanitarian Categorization Task baselines in Setting C with our suggested approach. We change the training data, which is listed in the columns, and correct the most recent catastrophe, which is the "California wildfires," as test data.

5.4 Ablation Study

In the ablation study, separate experiments were conducted for each component of the model illustrated in Figure 1. These components include textual multi-layer features, gated attention units (GAU), and low-rank fusion strategies. All experiments in this section were carried out under Setting A. The results in Table 5 reveal the significance of the GAU module, as the accuracy decreases from 91.21 to 89.42 when the GAU is eliminated. second, the choice to use multilevel embedding of text is justified: it can be seen experimentally that the accuracy drops to about 88 using the BERT final layer output replacing the multilevel feature joint representation. Third, by replacing the low rank-fusion of multi-modal representations with feature fusion[33] performance will drop significantly from 91.21 to 84.51. in terms of F1 scores, the macro F1 is reduced to 52.38 and the weighted F1 score is reduced to 82.56.

Table 5: Ablation Analysis of the Humanitarian Categorization Task in Setting A Using Our Suggested Method.

Model	Test Set		
	Accuracy	Macro F1	Weighted F1
MLLMF(Ours)	91.21	68.73	91.94
-Multi-layers features of text	88.26	55.76	86.54
-GAU	89.42	56.37	87.79
- LMF	84.51	52.38	82.56

6 Conclusion and Future Work

A novel multi-modal framework for combining text and visual inputs is presented in this research. The method introduces a new multi-level feature union module, which makes full use of features at different levels in the text as a basis for discrimination, enabling significant improvement in the effectiveness of the classification task even on small samples.

The framework also introduces strategies for joint gated attention units (GAUs) and low-rank fusion, emphasizing both intra-modal semantics and inter-modal associations of multimodal data. The effectiveness of this approach is assessed in two crisis tasks involving social media posts with images and textual captions. Experimental results demonstrate that the method not only surpasses mainstream multimodal combination approaches but also attains notable advantages in terms of model parameters and training speed. These advantages are crucial for the urgency of crisis detection tasks.

Although the present model achieves better results in terms of balancing accuracy and efficiency, however, the improvement in terms of accuracy is limited compared with the state-of-the-art model, on the one hand, because of the limitations of image and text feature extraction methods, improving them is expected to better obtain advanced feature representation of unimodal data, on the other hand, the heterogeneity of different modalities leads to the problem of modal imbalance, and subsequent work will investigate a resource balanced multimodal data classification method, which is an important work. Additionally, testing and expanding the approach to additional multimodal issues like sarcasm recognition in social media postings are planned for future study.

Acknowledgments:Funding Program of 20 Policies for Universities in Jinan, 202228120.

References

[1] Sun, Q.Y., Li, X.Y., Yu, F. Designing an emergency continuity plan for a megacity government: a conceptual framework for coping with natural catastrophes. International Journal of Critical Infrastructure Protection 2016, 13, 28–35.

[2] Huang, Q., Cervone, G. Usage of social media and cloud computing during natural hazards. In Cloud computing in ocean and atmospheric sciences; Elsevier, 2016; pp. 297–324.

[3] Puspita, I.A., Soesanto, R.P., Muhammad, F. Designing Mobile Geographic Information System for Disaster Management by Utilizing Wisdom of The Crowd. In Proceedings of the 2019 IEEE 6th International Conference on Industrial Engineering and Applications (ICIEA). IEEE, 2019, pp. 496–500.

[4] Amin, M.A., Ali, A.A., Rahman, A.M. Visual attention-based comparative study on disaster detection from social media images. Innovations in Systems and Software Engineering 2020, 16, 309–319.

[5] Tapia, A.G., Ramirez-Marquez, J.E. Understanding collective action through social media-based disaster data analytics. In Understanding Disaster Risk; Elsevier, 2021; pp. 297–318.

[6] Kemavuthanon, K., Uchida, O. Classiffcation of social media messages posted at the time of disaster. In Proceedings of the Information Technology in Disaster Risk Reduction: 4th IFIP TC 5 DCITDRR International Conference, ITDRR 2019, Kyiv, Ukraine, October 9–10, 2019, Revised Selected Papers 4. Springer, 2020, pp. 212–226.

[7] Acerbo, F.S., Rossi, C. Filtering informative tweets during emergencies: a machine learning approach. In Proceedings of the Conext Workshop on Ict Tools for Emergency Networks Disaster Relief, 2017.

[8] Neppalli, V.K., Caragea, C., Caragea, D. Deep neural networks versus naive bayes classiffers for identifying informative tweets during disasters. In Proceedings of the Proceedings of the 15th Annual Conference for Information Systems for Crisis Response and Management (ISCRAM), 2018.

[9] Alam, F., Offi, F., Imran, M. Crisismmd: Multimodal twitter datasets from natural disasters. In Proceedings of the Proceedings of the international AAAI conference on web and social media, 2018, Vol. 12.

[10] Schwartz, H.A., Giorgi, S., Kern, M.L., Park, G., Sap, M., Labarthe, D.R., Larson, E.E., Seligman, M., Ungar, L.H., et al. More evidence that Twitter language predicts heart disease: a response and replication 2018.

[11] Said, N., Ahmad, K., Riegler, M., Pogorelov, K., Hassan, L., Ahmad, N., Conci, N. Natural disasters detection in social media and satellite imagery: a survey. Multimedia Tools and Applications 2019, 78, 31267–31302.

[12] Madichetty, S., Sridevi, M. Detecting informative tweets during disaster using deep neural networks. In Proceedings of the 2019 11th International Conference on Communication Systems & Networks (COMSNETS). IEEE, 2019, pp. 709–713.

[13] Rudner, T.G., Rußwurm, M., Fil, J., Pelich, R., Bischke, B., Kopacková, V., Bilinski, P. Rapid Computer Vision-Aided Disaster Response via Fusion of Multiresolution, Multisensor, and Multitemporal Satellite Imagery. In Proceedings of the Proceedings of the First Workshop on AI for Social Good. Neural Information Processing Systems (NIPS-2018), Montreal, QC, Canada, 2018, pp. 3–8.

[14] Blevins, T., Kwiatkowski, R., Macbeth, J.C., McKeown, K., Patton, D., Rambow, O. Automatically processing tweets from gang-involved youth: towards detecting loss and aggression 2016.

[15] Shekhar, H., Setty, S. Disaster analysis through tweets. In Proceedings of the 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE, 2015, pp. 1719–1723.

[16] Sitaula, C., Basnet, A., Mainali, A., Shahi, T.B. Deep Learning-Based Methods for Sentiment Analysis on Nepali COVID-19-Related Tweets. Comput Intell Neurosci 2021.

[17] Li, X., Caragea, D., Zhang, H., Imran, M. Localizing and quantifying damage in social media images. In Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, 2018, pp. 194–201.

[18] Nalluru, G., Pandey, R., Purohit, H. Relevancy classification of multimodal social media streams for emergency services. In Proceedings of the 2019 IEEE International Conference on Smart Computing (SMARTCOMP). IEEE, 2019, pp. 121–125.

[19] Poria, S., Chaturvedi, I., Cambria, E., Hussain, A. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In Proceedings of the 2016 IEEE 16th international conference on data mining (ICDM). IEEE, 2016, pp. 439–448.

[20] Wörtwein, T., Scherer, S. What really matters—an information gain analysis of questions and reactions in automated PTSD screenings. In Proceedings of the 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE, 2017, pp. 15–20.

[21] Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M. Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847 2016.

[22] Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.P. Tensor fusion network for multimodal sentiment analysis. arXiv preprint arXiv:1707.07250 2017.

[23] Liu, Z., Shen, Y., Lakshminarasimhan, V.B., Liang, P.P., Zadeh, A., Morency, L.P. Efficient low-rank multimodal fusion with modality-specific factors. arXiv preprint arXiv:1806.00064 2018.

[24] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

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

[26] Hua, W., Dai, Z., Liu, H., Le, Q. Transformer quality in linear time. In Proceedings of the International Conference on Machine Learning. PMLR, 2022, pp. 9099–9117.

[27] Jawahar, G., Sagot, B., Seddah, D. What does BERT learn about the structure of language? In Proceedings of the ACL 2019-57th Annual Meeting of the Association for Computational Linguistics, 2019.

[28] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I. Attention is all you need. Advances in neural information processing systems 2017, 30.

[29] Dauphin, Y.N., Fan, A., Auli, M., Grangier, D. Language modeling with gated convolutional networks. In Proceedings of the International conference on machine learning. PMLR, 2017, pp. 933–941.

[30] Abavisani, M., Wu, L., Hu, S., Tetreault, J., Jaimes, A. Multimodal categorization of crisis events in social media. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 14679–14689.

[31] Kiela, D., Grave, E., Joulin, A., Mikolov, T. Efficient large-scale multi-modal classification. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2018, Vol. 32.

[32] Kiela, D., Bhooshan, S., Firooz, H., Perez, E., Testuggine, D. Supervised multimodal bitransformers for classifying images and text. arXiv preprint arXiv:1909.02950 2019.

[33] Ramachandram, D., Taylor, G.W. Deep multimodal learning: A survey on recent advances and trends. IEEE signal processing magazine 2017, 34, 96–108.