Focusing Attention across Multiple Images for Multimodal Event Detection

Yangyang Li
liyangyang@cetc.com.cn
National Engineering Laboratory
for Risk Perception and Prevention, caeit

Jun Li*
y.lijun@hotmail.com
University of Electronic Science
and Technology of China

Hao Jin
jinhao1@cetc.com.cn
National Engineering Laboratory
for Risk Perception and Prevention, caeit

Liang Peng
pliang951125@outlook.com
University of Electronic Science
and Technology of China

ABSTRACT

Multimodal social event detection has been attracting tremendous research attention in recent years, due to that it provides comprehensive and complementary understanding of social events and is important to public security and administration. Most existing works have been focusing on the fusion of multimodal information, especially for single image and text fusion. Such single image-text pair processing breaks the correlations between images of the same post and may affect the accuracy of event detection. In this work, we propose to focus attention across multiple images for multimodal event detection, which is also more reasonable for tweets with short text and multiple images. Towards this end, we elaborate a novel Multi-Image Focusing Network (MIFN) to connect text content with visual aspects in multiple images. Our MIFN consists of a feature extractor, a multi-focal network and an event classifier. The multi-focal network implements a focal attention across all the images, and fuses the most related regions with texts as multimodal representation. The event classifier finally predict the social event class based on the multimodal representations. To evaluate the effectiveness of our proposed approach, we conduct extensive experiments on a commonly-used disaster dataset. The experimental results demonstrate that, in both humanitarian event detection task and its variant of hurricane disaster, the proposed MIFN outperforms all the baselines. The ablation studies also exhibit the ability to filter the irrelevant regions across images which results in improving the accuracy of multimodal event detection.

CCS CONCEPTS

• **Information systems → Multimedia information systems**;
• **Computing methodologies → Computer vision tasks**.

*corresponding author

MMAsia '21, December 1–3, 2021, Gold Coast, Australia

ACM ISBN 978-1-4503-8607-4/21/12…$15.00
https://doi.org/10.1145/3469877.3495642

KEYWORDS

Multimodal Social Event Detection, Multi-focal Network, Focal Attention

ACM Reference Format:
Yangyang Li, Jun Li, Hao Jin, and Liang Peng. 2021. Focusing Attention across Multiple Images for Multimodal Event Detection. In *ACM Multimedia Asia (MMAsia '21), December 1–3, 2021, Gold Coast, Australia.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3469877.3495642

1 INTRODUCTION

In the past couple of years, with the popularization of the mobile Internet and the development of digital equipment, *e.g.*, smartphone, social media becomes more and more popular, and has been achieving tremendous progress. Benefiting from its speciality of convenience and propagation, people tend to record and post daily-life and other information on the social media, including texts, emojis, images, videos, *etc.* With these large-scale and various posts all over the world, it makes detecting events via social media information available in many areas [1, 14, 20]. Such social events detection is very important to a variety of real-world scenarios, such as, disaster situation updating, humanitarian planning and decision making, because it is somehow crowd-sourced and could be updated quickly. For example, during the torrential rain and flooding in Henan, July, 2021, people created an online working sheet that anyone could edit, and shared it on social media to report nearby citizens waiting for rescue[1]. This operation significantly improved the efficiency for saving lives.

Limited by the length of text content and processing techniques, early social event detection focuses on single-modal data, *e.g.*, texts, which could be formulated as topic modelling and tracking [3, 11, 13, 23, 24]. Since an image conveys more detailed aspects than a short text, visual texture and semantic learning are proposed to extract event clues from visual data, *e.g.*, images and videos, to leverage more detailed information [6, 7, 12, 25]. However, a post on social media usually consists of texts, images, audio as well as other meta-data, as illustrated in Figure 1. Analyzing only the texts or images can barely take fully advantage of multimodal information and may fail to capture the aspects we are interested in [21]. Researchers step further to simultaneously mine semantic clues from multimodal data, mainly from vision and language data, for social event detection [1, 17, 20]. We also mainly investigate the

[1]https://www.thepaper.cn/newsDetail_forward_13691244

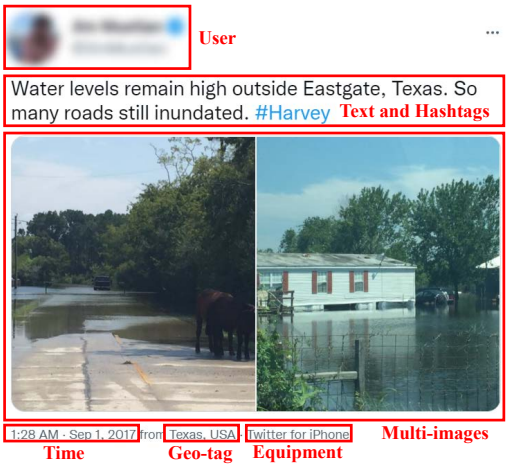

Figure 1: An example of multimodal social data. A post usually consists of contents of text and one or more images, as well as a variety of meta information, including user information, time stamp, geo-tag and *etc*.

multimodal social event detection in this work. The joint usage of data from multimodalities makes the model leverage the information more comprehensively and effectively, because the data from different modalities could supplement each other. At early stage, many simple ways, *e.g.*, sum and concatenation, are firstly introduced to fuse multimodal information. While such simple ways may introduce noise and decrease the effectiveness for multimodal event detection. Canonical correlation analysis (CCA) [9] was proposed to learn the joint semantic embedding to maximize the correlation between images and texts, and find more significant clues for social media event detection [10]. Motivated by the great success of attention mechanism [5, 19, 22], Abavisani *et al.* [1] propose a deep framework with cross-attention to avoid negative knowledge during fusion.

As everyone knows, a social media post may contain multiple images, *e.g.*, up to four images for Twitter[2] and nine images for Weibo[3]. Nevertheless, existing methods can only handle one image for each sample, integrating with text representation or other information. Ofli *et al.* [18] split all the samples with multiple images as training, and all the test samples are with only one image. The authors then repeatedly pair the text with different images in a post and train the model with an image-text pair. These methods obviously would learn incorrect correlations between the text and image and result in wrong prediction, because not all the images in a post are related to its topics.

To address aforementioned issues for multimodal social event detection, in this paper we propose a novel Multi-Image Focusing Network, referring as MIFN, to handle multiple images in a post. Specifically, we first employ a bidirectional GRU and pre-trained Faster-RCNN as the feature extractor, to learn the text representations and image region representations. To align the semantic concepts in text and visual aspects in images, inspired by [15], we devise a multi-focal network for MIFN, focusing attention on the regions most related to query text contents. This operation across multiple images not only filters the irrelevant regions in the same image, but also avoids the ones across images, even totally ignores the irrelevant images by filtering all the regions. The relevant visual aspects are then integrated with semantic contents by weighted sum, to obtain the multimodal event representation. Finally, we make the prediction based on previous multimodal representations utilizing an event classifier, which first pools the multimodal representation and predicts the label with softmax operation. To verify the effectiveness of proposed approach, we attempt to classify the natural disasters via the relevant multimodal posts on social media, and conduct extensive experiments on CrisisiMMD dataset. The experimental results demonstrate that our method focusing attention across multiple images could significantly improve the prediction accuracy.

In summary, the main contributions of this work are as below:

- We propose a Multi-Image Focusing Network for multimodal social event detection. To the best of our knowledge, we are one of the first to simultaneously integrate multiple images and text content of a social media post for multimodal event understanding and detection.
- We design a multi-focal network to focus attention across multiple images, which makes the model could fuse text and vision from different images.
- We conduct extensive experiments on CrisisMMD dataset and the experimental results demonstrate the effectiveness of proposed approach.

2 APPROACH

The framework of our proposed **MIFN** is shown in Figure 2. Our model mainly consists of three components: feature extractors, multi-focal network and event classifier. For a given tweet, consisting of a text and multiple images, we first use feature extractors to obtain the feature maps of the text and images. Second, we leverage the multi-focal network to initialize the weight of different image regions related to a given word in tweet text, and filter out the irrelevant image regions. We further fuse tweet text and relevant image regions by the attention module as the final tweet embedding. Finally, the final tweet embedding is fed into the event classifier, composed of an avg-pooling layer, a full connection layer and a softmax layer, to predict the event category.

2.1 Feature Extractors

A multimodal tweet contains the main content of text sentence(s) s and a set of images V. As the text and images are all related to the same event, the object of s describes may distribute among multiple images in $V = \{v_i\}_{i=1}^{N}$, where v_i is i-th image, N is the number of all corresponding images. Therefore, different from extracting the feature of the whole image, we employ the pretrained Faster R-CNN with ResNet-101 [4] to detect objects in an image and extract their features. For a given image v_i, each object is marked by a rectangular box, which named image region, thus the visual embedding of v_i is $\{v_{ij}\}_{j=1}^{M} \in \mathrm{R}^{d_v \times M}$, where v_{ij} is the feature of j-th image region. Finally, the image set V is embedded as $V = \{\{v_{1j}\}_{j=1}^{M}; \cdots; \{v_{Nj}\}_{j=1}^{M}\} \in \mathrm{R}^{d_v \times (M+N)}$.

[2]https://twitter.com
[3]https://weibo.com

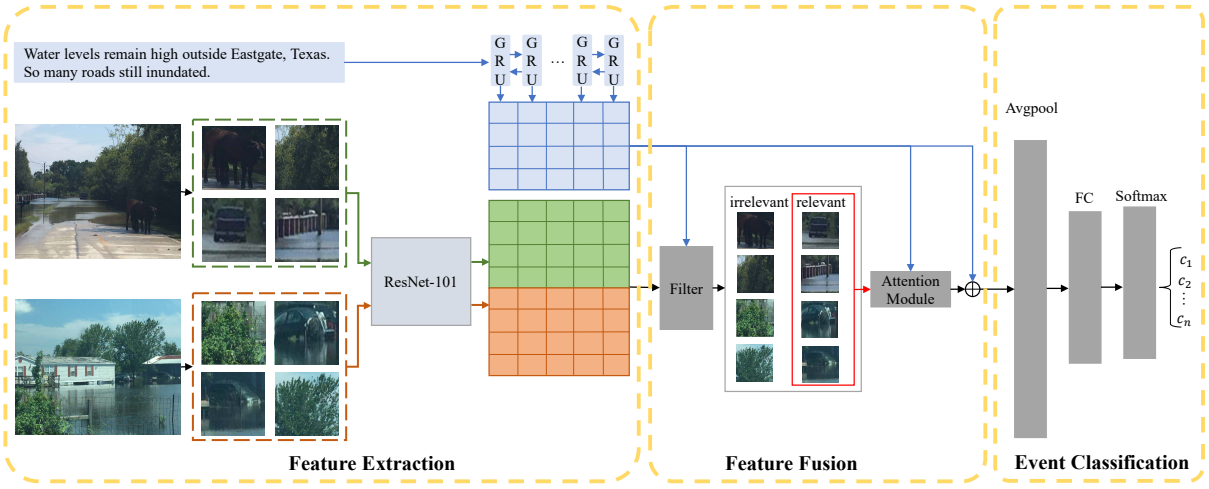

Figure 2: The framework of the proposed MIFN. Given a tweet, consists of a text and multiple images, we extract the textual feature map and visual feature map. For each word, we further filter out the irrelevant image regions based on the word embedding, and fuse the word embedding with relevant images regions. The final tweet embedding, composed of textural embedding and corresponding relevant visual embedding, is fed into event classifier.

To encode the semantic information of words in s, we first use the word embedding matrix to embed the word sequence. Then, the embedded word sequence is input to a bidirectional recurrent neural network, *e.g.*, Bi-GRU, to obtain the hidden sequence $H_L = \{h_t^L\}_{t=1}^T \in \mathrm{R}^{d_h \times T}$ and $H_R = \{h_t^R\}_{t=1}^T \in \mathrm{R}^{d_h \times T}$, where d_h is the dimension of the hidden layer and T is the length of the text sequence s. Finally, to fuse the bidirectional semantic information of a given word, we combine the word embedding by averaging the hidden vectors in the two direction:

$$e_t = (h_t^L + h_t^R)/2; \tag{1}$$

where e_t is the word embedding of t-th vector. The textual embedding matrix of s is $E = \{e_t\}_{t=1}^T \in \mathrm{R}^{d_h \times T}$.

2.2 Multi-focal Network

Previous multimodal event detection works focused on the the pair of text and single image, but ignored that the tuple of text and multiple images is more common in realistic scenario, as mentioned in Figure 1. As the limited words are allowed on social media, users could share multiple images as the supplementary information of what the text described. However, we find out the image regions, that the object of this text describes corresponds to, may not concentrate on a single image. To better fuse the textual feature and visual feature, we select all the relevant regions of images. The final tweet embedding is composed of the textual embedding and the visual embedding of relevant regions.

Specifically, we first initialize the relevant score matrix R of the words in s toward all the regions in V:

$$\begin{aligned} Q &= W_Q E \quad C = W_C V; \\ R &= \mathbf{softmax}(Q^T C); \end{aligned} \tag{2}$$

where $W_Q \in \mathrm{R}^{d_i \times d_h}; W_C \in \mathrm{R}^{d_i \times d_v}$ are transform matrix, $R = \{r_i\}_{i=1}^T$ and $r_i = \{r_{ij}\}_{j=1}^{M+N} \in \mathrm{R}^{1 \times (M+N)}$.

For each word, we then assume that relevant regions obtain the higher relevant score than irrelevant regions [15]. Therefore, we filter out the regions with the relevant score lower than the average score of all the regions. The function of filtering is formulated as:

$$f(r_{ij}) = \begin{cases} 1 & \text{if} \quad r_{ij} > \frac{1}{M+N}; \\ 0 & \text{others}; \end{cases} \tag{3}$$

where r_{ij} is the relevant score of i-th word and j-th image region, and $j \in [0; M+N-1]$.

Finally, we reassign the weight for the rest of relevant regions and define the weight as:

$$\hat{r_{ij}} = \frac{r_{ij} f(r_{ij})}{\sum_{j=0}^{M+N-1} r_{ij} f(r_{ij})}; \tag{4}$$

For the i-th word, the reassigned weight vector is $\hat{r_i} = \{\hat{r_{ij}}\}_{j=1}^{M+N} \in \mathrm{R}^{1 \times (M+N)}$. As for the text sequence s, the reassigned weight matrix is $\hat{R} = \{\hat{r_i}\}_{i=1}^T \in \mathrm{R}^{T \times (M+N)}$. After that, we leverage attention module to fuse the text-related visual feature and the textual feature. The attention module is defined as:

$$M = Q + C\hat{R}^T; \tag{5}$$

where $M \in \mathrm{R}^{d_i \times T}$ is the feature fusion map.

2.3 Event Classifier

To retain all the information of M, we employ the avg-pooling to obtain the final tweet embedding:

$$m = \mathbf{avg\text{-}pooling}(M); \tag{6}$$

where $m \in \mathrm{R}^{d_i \times 1}$. Then, we acquire the possibility of each event category:

$$\hat{y} = \mathbf{softmax}(Wm + b); \tag{7}$$

where W and b are learnable parameters, and $\hat{y}$ is the probability vector of all the event category. Finally, we use the cross-entropy

Table 1: The data statistics of Task 2 (Humanitarian Categories).

Event Name	Train Set	Dev Set	Test Set
Infrastructure and utility damage	458	98	99
Rescue volunteering or donation effort	726	156	156
Affected individuals	61	13	14
Not-humanitarian	2,650	568	568
Other relevant	1,164	250	250
Total	5,059	1,085	1,087

Table 2: The data statistics of Hurricane Disaster.

Event Name	Train Set	Dev Set	Test Set
Infrastructure and utility damage	356	76	77
Rescue volunteering or donation effort	548	118	118
Affected individuals	26	6	6
Not-humanitarian	1,834	393	394
Other relevant	1,017	218	219
Total	3,781	811	814

loss as the objective function to optimize all the parameters during training.

3 EXPERIMENTS

3.1 Dataset and Setting

There are few public multimodal event detection datasets, thus we conduct experiments on one public dataset, **CrisisMMD** [2], with several prediction tasks.

CrisisMMD is a natural disaster event dataset, which is collected during seven natural disasters, and annotated with the text-image pair for three tasks: (1) Informative vs. Not Informative, (2) Humanitarian Categories, (3) Damage Severity. Among these tasks, Task 1 (Informative vs. Not Informative) is to determine whether the tweet is useful for humanitarian aid purposes and Task 3 (Damage Severity) is applied only on images, hence only Task 2, *i.e.*, Humanitarian Categories, is fit for our multimodal event detection. Following the setting of [18], we merge the eight categories into five humanitarian categories.

In this paper, we study the event detection of single text and multiple images. To construct the text-images pair, we merge the images which correspond to the same tweet id. In addition, Task 2 is about the humanitarian event during different natural disasters, which contains hurricane, wildfires, earthquake and floods. To avoid the different disasters effect image background and semantics of tweet text, we design another experiment task that detecting humanitarian event about a specific disaster, hurricane which has more samples.

Table 3: The results of Task2 (Humanitarian Categories).

Task2 (Humanitarian Categories)			
Model	**Acc**	**Macro F1**	**Weighted F1**
Bi-GRU	74.5	55.0	73.6
ResNet-101	78.7	60.1	78.0
GRU+ResNet-101			
adding	82.2	67.5	81.8
max-pooling	78.2	60.0	77.7
avg-pooling	82.9	63.5	81.4
MIFN	**83.8**	**74.7**	**83.6**

Finally, for each event category, we randomly choose 70% data for training, 15% for development and 15% for testing. The statistics of these two tasks are shown in Table 1 and Table 2.

3.2 Baselines

To evaluate the effectiveness of our proposed model, we design two groups of comparative experiments.

The unimodal model only leverages the unimodal, *i.e.*, text or images, as the representation of a given tweet.

- **Bi-GRU**: The textual processor. Bi-GRU extracts the semantic information of the given text by bi-direction process.
- **ResNet-101**: The visual processor. ResNet-101 [8] has showed a great power on image classification. We use the ResNet-101 pretrained on ImageNet to extract the visual feature.

The multimodal model attempts to detect events based on employing more information. The challenge is that finding out a better method to fuse the textual and multiple image representations. To compare with our model, we obtain the textual embedding and visual embedding by Bi-GRU and pre-trained ResNet-101 respectively and fuse these two embedding by max-pooling, average-pooling or adding.

3.3 Implementation Details

We implement all the models with Pytorch. For baselines, we set the hidden size of Bi-GRU to 1024 and use the pre-trained ResNet-101 in torchvision. About our model, $d_h; d_v; d_i$ are 1024, 2048, 1024 respectively. The number of regions M is set to 36. For the images which have less regions, we pad them to 36. The batch size, optimizer and L2 regularization weight are set to 32, Adam and 0.0001 for all the models. We initialize learning rate to 0.01 and reduce it with the factor (0.1) every 15 epochs.

3.4 Overall Performance

The comparison results of our MIFN and baselines on Humanitarian Categories and Hurricane Disaster are shown in Table 3 and Table 4.

First, among the unimodal methods, ResNet-101 (pure images) is better than GRU (pure text). In twitter, users are allowed to share messages of 140 characters or less [16]. It is difficult to describe the detail of damage with less words during the natural disaster. However, the images could provide more intuitive information about the disaster scene. This is the reason that leveraging the

Table 4: The results of Hurricane Disaster.

Hurricane Disaster			
Model	**Acc**	**Macro F1**	**Weighted F1**
Bi-GRU	73.8	55.4	73.2
ResNet-101	78.6	61.4	78.5
GRU+ResNet-101			
adding	79.2	62.4	78.9
max-pooling	76.9	58.9	76.5
avg-pooling	80.2	62.8	79.8
MIFN	**84.3**	**66.1**	**83.9**

visual information has the better performance than the textual information.

Second, about multimodal methods, we observe that different fusion methods would result in the performance varies considerably. For textual feature extracted by Bi-GRU and visual feature extracted by ResNet-101, we fuse these multimodal feature by max-pooling, average-pooling, or directly adding. The max-pooling gets the worse performance which is even worse than the unimodal method. This may be that the max-pooling removes some information in the single modality by simple way, which would expand the effect of the irrelevant information. However, the average-pooling and the adding method retain the information of both modalities and achieve better performance.

Third, compared with the multimodal methods, our proposed method (MIFN) gains 1.6%-5.6% improvement on Task 2 and 4.1%-7.4% improvement on our designed task. The multimodal baselines regard that all the regions in images are related to the tweet text, and fuse the features of different modalities in a simple way. However, our model filters out the irrelevant image regions of corresponding words before fusing the features. By this way, each word could be fused with the semantic-related image regions and a better multimodal tweet embedding is obtained. This results in that MIFN reaches the best performance among all the baselines. In addition, we observe that more improvement is gained on Hurricane Disaster. In the same disaster, the texts and images have the more same semantic. It may be more easier for multi-focal network to select the relevant image regions.

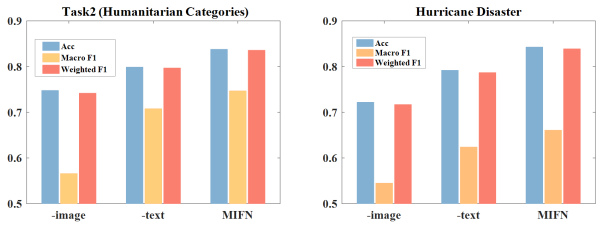

Figure 3: The performance of different variants of our proposed MIFN on humanitarian event detection. -image/-text represents the MIFN variant without visual information/textual information.

Table 5: The ablative analysis of MIFN on Task2 (Humanitarian Categories). -filter function represents the MIFN variant without filter function; max-pooling represents that MIFN variant leverage the avg-pooling to replace max-pooling in event classifier; single-image means the MIFN variant processes the text-image pair.

Task2 (Humanitarian Categories)			
Model	**Acc**	**Macro F1**	**Weighted F1**
MIFN	**83.8**	**74.7**	**83.6**
-filter function	82.9	71.3	82.5
max-pooling	81.8	63.5	81.2
single-image	82.9	64.1	82.5

Table 6: The ablative analysis of MIFN on Hurricane Disaster. -filter function, max-pooling and single-image represents the same MIFN variants with Table 5.

Hurricane Disaster			
Model	**Acc**	**Macro F1**	**Weighted F1**
MIFN	**84.3**	**66.1**	**83.9**
-filter function	80.1	62.9	79.7
max-pooling	81.3	63.1	81.1
single image	83.7	65.8	83.3

3.5 Ablation Study

To investigate the effectiveness of textual and visual information, we design two variants of MIFN, -text and -image. As shown in Figure 3, without either of these two information, the performance drops. This indicates that textual and visual information are both important to understand the event. In particular, the performance of the variant without visual information drops significantly, which is consistent with the observation analysed in Section 3.4.

In addition, we devise another three variants to analyze the effect of the multiple images, filter function and pooling method. The results are shown in Table 5 and Table 6.

First, different from most multimodal event detection works, we study the multimodal event detection under a more realistic scenario, which has the combination of single text and multiple images. Therefore, to investigate the effect of multi-images, we select the first image in each multimodal tweet and extract its feature map as the visual information. We find that more images provide more related information for event detection.

Second, the filter function is the main component of the multi-focal network. It filters out the irrelevant image regions of a given word and this is the key point that we argue in this paper. To analyze the affect of filter function in fusing multimodal features, we remove the filter function and directly apply the attention module to fuse the textual and visual features. By this way, all the regions in multiple images are more or less relevant to the given word. Although the model would reduce the relevant scores of irrelevant regions by iterative learning, the scores would not be reduced to zero and irrelevant regions still affect the result of classification. The result of Hurricane Disaster drops significantly. This is because the images

in same disaster may have more same semantic information and it is more need the filter function to select the salient regions.

Third, we leverage the avg-pooling to obtain the final tweet embedding in MIFN. To explore the effect of different pooling method, we replace the avg-pooling with max-pooling. The result shows that max-pooling gets the worse performance. This indicates that avg-pooling retains all the information contained in the fusion feature map and more information results in the improvement of classification performance.

4 CONCLUSION

In this paper, we studied the multimodal event detection under a more realistic scenario, which has the pair of single text and multiple images. Further, we proposed a Multi-Image Focusing Network (MIFN) to fuse the textual feature and visual feature. For multiple images, we leveraged the pretrained Faster R-CNN to obtain the image regions of each image and extract their features. Different from fusion method of most multimodal event detection, we filtered out the irrelevant image regions towards a given word before fusing multimodal information. To explore the effectiveness of fusing multiple images and the proposed filter function, we designed two variants of MIFN. The experimental results show that MIFN is better than these two variants, which demonstrate the advantage of using multiple images to expand textual information, and filtering out irrelevant information before multimodal fusion.

ACKNOWLEDGMENTS

This work was supported in part by Science and Technology Development Fund of CETC.

REFERENCES

[1] Mahdi Abavisani, Liwei Wu, Shengli Hu, Joel Tetreault, and Alejandro Jaimes. 2020. Multimodal categorization of crisis events in social media. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14679–14689.

[2] Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018. Crisismmd: Multimodal twitter datasets from natural disasters. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.

[3] James Allan. 2002. Introduction to topic detection and tracking. In *Topic detection and tracking*. Springer, 1–16.

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6077–6086.

[5] Yi Bin, Yang Yang, Fumin Shen, Ning Xie, Heng Tao Shen, and Xuelong Li. 2018. Describing video with attention-based bidirectional LSTM. *IEEE transactions on cybernetics* 49, 7 (2018), 2631–2641.

[6] Yi Bin, Yang Yang, Fumin Shen, and Xing Xu. 2016. Combining multi-representation for multimedia event detection using co-training. *Neurocomputing* 217 (2016), 11–18.

[7] Xiaojun Chang, Zhigang Ma, Yi Yang, Zhiqiang Zeng, and Alexander G Hauptmann. 2016. Bi-level semantic representation analysis for multimedia event detection. *IEEE transactions on cybernetics* 47, 5 (2016), 1180–1197.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[9] Harold Hotelling. 1992. Relations between two sets of variates. In *Breakthroughs in statistics*. Springer, 162–190.

[10] Po-Yao Huang, Junwei Liang, Jean-Baptiste Lamare, and Alexander G Hauptmann. 2018. Multimodal filtering of social media for temporal monitoring and event analysis. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*. 450–457.

[11] Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation* (1972).

[12] Zhen-zhong Lan, Lei Bao, Shoou-I Yu, Wei Liu, and Alexander G Hauptmann. 2012. Double fusion for multimedia event detection. In *International Conference on Multimedia Modeling*. Springer, 173–185.

[13] Nut Limsopatham and Nigel Collier. 2016. Bidirectional LSTM for named entity recognition in Twitter messages. (2016).

[14] Zhihong Lin, Huidong Jin, Bella Robinson, and Xunguo Lin. 2016. Towards an accurate social media disaster event detection system based on deep learning and semantic representation. In *Proceedings of the 14th Australasian Data Mining Conference, Canberra, Australia*. 6–8.

[15] Chunxiao Liu, Zhendong Mao, An-An Liu, Tianzhu Zhang, Bin Wang, and Yongdong Zhang. 2019. Focus your attention: A bidirectional focal attention network for image-text matching. In *Proceedings of the 27th ACM International Conference on Multimedia*. 3–11.

[16] Kristen Lovejoy, Richard D Waters, and Gregory D Saxton. 2012. Engaging stakeholders through Twitter: How nonprofit organizations are getting more out of 140 characters or less. *Public relations review* 38, 2 (2012), 313–318.

[17] Andrew J McMinn, Yashar Moshfeghi, and Joemon M Jose. 2013. Building a large-scale corpus for evaluating event detection on twitter. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 409–418.

[18] Ferda Ofli, Firoj Alam, and Muhammad Imran. 2020. Analysis of social media data using multimodal deep learning for disaster response. *arXiv preprint arXiv:2004.11838* (2020).

[19] L. Peng, Y. Yang, Z. Wang, Z. Huang, and H. T. Shen. 2020. MRA-Net: Improving VQA via Multi-modal Relation Attention Network. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020), 10.1109/TPAMI.2020.3004830.

[20] Georgios Petkos, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2012. Social event detection using multimodal clustering and integrating supervisory signals. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*. 1–8.

[21] H. T. Shen, Y. Zhu, W. Zheng, and X. Zhu. 2020. Half-Quadratic Minimization for Unsupervised Feature Selection on Incomplete Data. *IEEE Transactions on Neural Networks and Learning Systems* (2020), 10.1109/TNNLS.2020.3009632.

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

[23] Kang Xu, Guilin Qi, Junheng Huang, Tianxing Wu, and Xuefeng Fu. 2018. Detecting bursts in sentiment-aware topics from social media. *Knowledge-Based Systems* 141 (2018), 44–54.

[24] Xiao Yang, Craig Macdonald, and Iadh Ounis. 2018. Using word embeddings in twitter election classification. *Information Retrieval Journal* 21, 2 (2018), 183–207.

[25] Yang Yang, Jie Zhou, Jiangbo Ai, Yi Bin, Alan Hanjalic, Heng Tao Shen, and Yanli Ji. 2018. Video captioning by adversarial LSTM. *IEEE Transactions on Image Processing* 27, 11 (2018), 5600–5611.