




Causal Data Fusion for Multi-modal Disaster Classification in Social Media

Raha Moraffah^(✉), Suraj Jyothi Unni, Adrienne Raglin, and Huan Liu

¹ Computer Science & Engineering, Arizona State University, Tempe, AZ, USA
{rmoraffa, sjyothiu, huanliu}@asu.edu

² Army Research Lab, USA adrienne.raglin2.civ@mail.mil

Abstract. Analyzing social media data for disaster assessment and relief has gained tremendous attentions. Recent studies have recognized the importance of using multi-modal social media data (i.e., both text and image) for disaster relief. We propose a novel multi-modal causal data fusion mechanism, which models the relationships between modalities via causal graphs. For each causal graph, our framework proposes an approach to combine the modalities such that the integrated data only contain informative, non-redundant information for classification. We further propose to eliminate the need for relationships between modalities (thus the causal graphs) to be known a priori. Our proposed ensemble-based approach considers relations for each input, performs data integration accordingly, and draws the final prediction via voting. Our experiments on a widely-used real-world dataset demonstrate that our proposed method outperforms existing baselines in terms of classification performance on three disaster-related tasks.

Keywords: Multi-modal data fusion · Multi-modal classification · Disaster relief · Causality · Social media analysis · CrisisMMD.

1 Introduction

With the rapid growth of social media usage, people often share information in the forms of texts and images on social media platforms. During natural disasters, critical information such as reports of injured or deceased people, missing individuals, infrastructure damages and severity of damages is posted on social media. Analyzing such information can help the humanitarian organizations and governments make critical decisions in timely manner. Therefore many machine learning frameworks have been developed to process social media data and perform decision making for disaster relief [10,9,19,25].

Most existing frameworks for disaster relief are designed for uni-modal data such as either text [10,17], or image [16,4]. However, social media data are often multi-modal by nature, i.e., contain both image and text. Recently, utilizing multi-modal data for disaster relief has attracted much attentions [25,1,19] due to their richer information. To deal with multi-modal data, researchers have developed multi-modal machine learning frameworks to process information from

multiple modalities, integrate them into one unified source of information and feed that source to a classifier for decision making. Existing multi-modal frameworks often discard the relationships between them when concatenating the modalities to fuse data. However, it has been shown that data modalities, i.e., text and image, can have different relations [21]. In particular, two modalities can have complementary relationship (i.e., provide complementary information on the same event), similarity relationship (i.e., contain same or similar information), or unrelated relationship (i.e., do not share any information and are independent of each other). We argue that it is necessary to consider these relationships while combining these modalities. For example, in Figure 1a text and the image provide complementary information as they both contain information about a school collapsed by the earthquake. The text discusses the number of casualties and the image shows the ongoing relief effort. Here, both text and image provide complementary informative pieces of information as well as repetitive information since they both discuss the same event. Thus the data fusion approach needs to be able to extract all relevant complementary information from both modalities while excluding redundant information. In Figure 1b, both text and the image provide similar information regarding the same event - a tree that fell on a house. Since both modalities in this case provide similar information, simply concatenating them without considering their relationship results in redundant information. In Figure 1c, the text and image provide independent information about hurricane. While the text discusses the residents frustration, the image shows the infrastructures and buildings. In this case, all information in both modalities need to be utilized for classification.

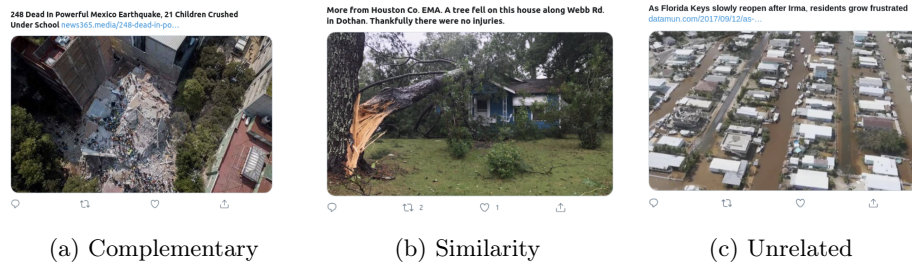


Fig. 1: Different types of relationships between modalities. Figures are from [21].

It is worth noting that irrelevant or redundant information could impact the prediction negatively or increase the possibility of wrong decision [8]. Therefore, it is necessary for multi-modal data fusion mechanisms to take modalities relationships into account and integrate different modalities accordingly. To achieve this, we propose a causal multi-modal data fusion approach that considers the relationships between the modalities and combine them based on their relations. We perform extensive experiments on the widely used real-world disaster-related dataset, CrisisMMD [3] and report the results on three tasks, namely: Infor-

mativeness; Humanitarian Event Classification; and Severity of damages. The results demonstrate that Causal-MMC outperforms existing baselines, including both uni-modal and multi-modal frameworks, in terms of classification metrics.

2 Related Work

Early research on disaster relief mainly utilized uni-modal data, i.e., only text or image datasets. Text-based approaches leverage textual data to detect crisis events. Kumar et al. [13] propose TweetTracker to track, analyze, and monitor text tweets immediately after a disaster. Bai et al. [5] use word representations to classify disaster-related messages on Weibo. Nguyen et al. [17] propose a convolutional neural network to classify tweet texts. Similarly, to analyze image datasets, frameworks such as CNNs [14], de-duplication and relevancy filtering mechanisms [18] have been proposed. Recent studies have shed lights on the importance of using multi-modal data (i.e., both text and image) to make more accurate and robust predictions for disaster relief. These frameworks often propose a data fusion mechanism to combine the text and image information effectively. For example, Ofli et al. [19] and Zou et al. [25] propose to transform both modalities to a fixed low dimensional space via fully-connected layers and concatenate them. Abavisani et al. [1] propose a cross attention module to filter misleading components from weak modalities. Sun et al. [22] improves the attention-based frameworks by adding a gating module. However, none of the aforementioned frameworks directly consider the relationships between text and image modalities. Moreover, attention-based frameworks implicitly capture the correlations between modalities which makes them prone to biases [15]. Our framework, on the other hand, proposes a causal data fusion approach to reflect the causal relations between modalities while performing data integration. Causal data integration for multi-modal data is underexplored. Körding et al. [11] propose a causal data integration approach for sensors. Wu et al. [24] propose to measure the causal strength of each modality of the outcome. None of these works are proposed for multi-modal disaster classification.

3 Proposed Framework

In this section, we propose our novel causal data fusion mechanism. Since the causal data fusion is designed for multi-modal classification, we embed it within a multi-modal classification pipeline and refer to the entire framework as Causal Multi-Modal Classifier (Causal-MMC). Causal-MMC takes image-text pairs as inputs and outputs their classification labels. The proposed Causal-MMC is equipped with a novel causal data fusion mechanism, which models the relationships between the modalities with causal graphs and combines them accordingly. The integrated data is then fed to a classifier to make predictions. In particular Causal-MMC consists of four components: (1) Image feature extractor: which is in charge of extracting image feature representations from raw images; (2) Text embedding extractor: which is in charge of extracting text embeddings from raw

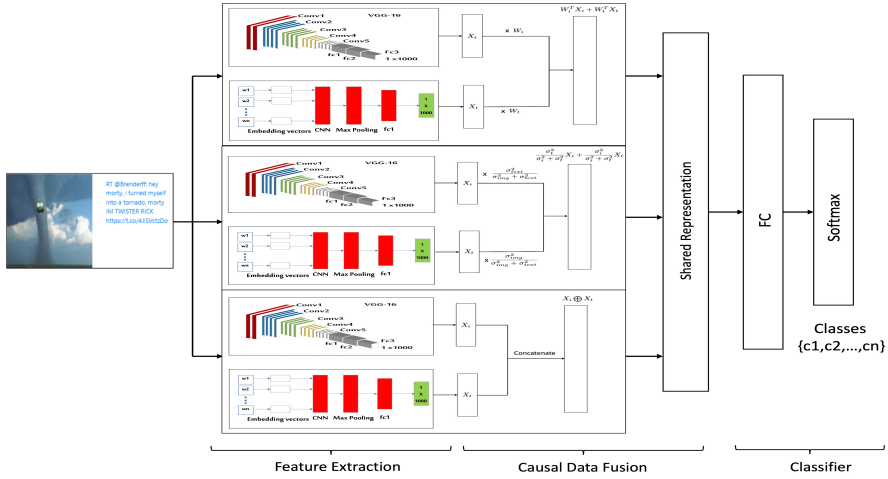


Fig. 2: An overview of Causal-MMC.

textual data; (3) Causal multi-modal data fusion module: our proposed data fusion method which combines the extracted image and text features from step (1) and (2) based on their causal relationships; and (4) Classifier: which takes the integrated unified embedding (generated in step (3)) as input and outputs its classification labels. An overview of Causal-MMC is shown in Figure 2. In the following we explain each part of the model in detail:

3.1 Image Feature Extractor

We use Convolutional Neural Network (CNN) to extract the feature representations from the raw image. In our model, we adopt VGG-16 [20], which has been widely used in previous disaster-relief research and has demonstrated impressive performance [19,25]. To speed up the model convergence and enhance its performance, based on the idea of transfer learning [23], we initialize the weights of VGG-16 with pretrained weights on ImageNet dataset [6] and fine-tune the model on our CrisisMMD dataset. After the fine-tuning step, we extract the features from the last layer before the fully-connected classification layer (1000 dimensions) and use it as feature representations.

3.2 Text Embedding Extractor

Despite their simple architectures, convolutional Neural Networks (CNNs) have demonstrated good performance for disaster text classification tasks [19,17]. Therefore, for the sake of simplicity, we also utilize a CNN-based architecture to extract text features from raw textual data, although it can be replaced with any state-of-the-art language models such as BERT [7]. Particularly, we use the architecture proposed in [19], which consists of a series of convolutional and

fully-connected layers with rectified linear units (ReLU) [12] as the nonlinear activation function. The input to the CNN is provided via word2vec model pre-trained on a disaster-related dataset of size 364 million tweets as discussed in [2]. The last fully connected layer of 1000 dimensions is used as feature representation for the text data.

3.3 Novel Multi-modal Causal Data Fusion Module

In this section, we explain our novel causal data fusion module, which takes multi-modal data (i.e., text and image) as inputs, models the causal relationships between the modalities, combines the knowledge extracted from each modality based on their causal relations to avoid irrelevant and redundant information, and outputs one unified integrated data source which contains predictive information from both modalities and can be used for classification. Previous

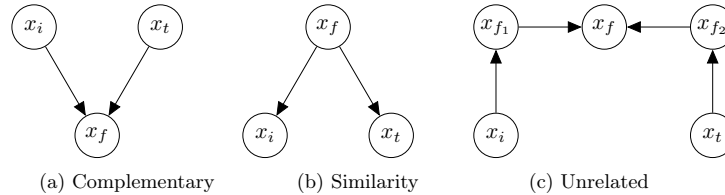


Fig. 3: Causal graphs used to model each type of modalities relationship.

study [21] shows that text and image data can have three different types of relationships: (1) Complementary relationship: where both text and image contain predictive complementary information about the same event; (2) Similarity relationship: text and image provide similar or same information on an event; and (3) Unrelated: text and image contain completely independent pieces information and are generated from different sources. For this case, each may provide different but useful information for reasoning. We propose to model these three types of relationships between modalities via three different causal graphs and propose an approach to integrate these modalities for each case:

Case 1: Complementary relationship For this case text and image contain complementary predictive information about the same event used to do classification. We therefore need to consider both modalities in generating the integrated data. We model this type of relationship with the causal graph shown in Figure 3a. As shown in the figure, both features extracted from text and image modalities contribute to the final data source (X_f). To integrate data modalities that follow this model, we propose the final source of data to be a function of both modalities. While this function can be parameterized with non-linear transformations, to maintain the explainability and simplicity of the model, we model the integrated data as a linear combination of both modalities:

$$X_f = W_i^T X_i + W_t^T X_t, \quad (1)$$

where X_i and X_t are features extracted from Image Feature Extractor(Sec. 3.1) and Text Embedding Extractor(Sec. 3.2), respectively. W_i and W_t are learnable weights, which are learned during training of the classifier.

Case 2: Similarity relationship For this case, both modalities contain same or similar information about the same event, thus can be considered different views of the same event. Here, if both modalities are considered independent and are combined for example via concatenation, the combined data may contain redundant information which can cause degradation in the classifier’s performance [8]. To obtain optimal predictive feature representations from the modalities while minimizing the redundancy, we propose to model their relationship with the causal graph presented in Figure 3b. In this causal graph, features extracted from the two modalities (i.e., X_i and X_t) are both children of the same event (X_f) and can be considered as noisy views of that event. X_f contains the optimal predictive set of features required for the classification without redundant information. To obtain X_f from X_i and X_t , we formulate the generation of X_i and X_t with $X_i = X_f + \epsilon_i$ and $X_t = X_f + \epsilon_t$, where ϵ_i and ϵ_t are two Gaussian noises. These generation models essentially imply that X_i and X_t are two noisy views of X_f , which are generated by adding Gaussian noises to it. Given the generation equations, we can model the $p(X_i | X_f) = \mathcal{N}(X_f, \sigma_i^2)$ (we can write a similar equation for X_t). Using Bayes rule we have: $p(X_f | X_i, X_t) \propto p(X_f | X_i) \times p(X_f | X_t)$. Applying Maximum Likelihood, we get:

$$X_f = \frac{\sigma_t^2}{\sigma_i^2 + \sigma_t^2} X_i + \frac{\sigma_i^2}{\sigma_i^2 + \sigma_t^2} X_t, \quad (2)$$

where σ_i^2 and σ_t^2 are variances of Gaussian noises ϵ_i and ϵ_t . In our experiments we consider the ratios $\lambda_i = \frac{\sigma_t^2}{\sigma_i^2 + \sigma_t^2}$ and $\lambda_t = \frac{\sigma_i^2}{\sigma_i^2 + \sigma_t^2}$ as hyperparameters and select the optimal value via grid search in the range $\{0.1, 0.2, \dots, 0.9\}$.

Case 3: Unrelated For this case modalities do not share any information. They contain independent pieces of information about the final data. We propose to model this relationship with the graph proposed in Figure 3c. In this graph, each modality is generated from a different source of data (x_{f_1} and x_{f_2}), therefore are independent of each other. Both X_{f_1} and X_{f_2} independently contribute to the final integrated data X_f . To get the final unified source of data for classification, we propose to integrate the two modalities via concatenation. Basically, the final source used for the classification is the concatenation of features extracted from both text and image modalities (i.e., X_t and X_i).

Causal Data Fusion Mechanism without Knowledge of Modalities Relationships Despite successfully modelling the relationships between modalities, our approach faces some challenges in real world implementation. Particularly, in real-world datasets, the relationships between the modalities or their corresponding causal graphs are often not given. Therefore, we are not able to determine which causal graph can be used to combine the modalities and generate the integrated data for inputs. One possible solution is to manually provide the relations for all data samples in the dataset which is extremely labor-intensive and time-consuming especially for the big datasets we work with. To address this problem, we propose an ensemble-based approach. Our approach considers all three types of relations (causal graphs) for each sample, generate the integrated data for each case and performs classification on the integrated data. The final decision is decided via voting on three classification results predicted for each integrated data source. This is as if we have an ensemble classifier, which consists of three classifiers with three different data integration approaches accounting for different types of relations between modalities.

3.4 Classifier

As explained in Sec. 3.3, for each data input, three integrated data representations are generated (one for each type of relationship). Each integrated representation is then fed to a fully-connected layer followed by a soft-max layer to provide class confidence scores for classification. The class with highest confidence score is selected as the final label for each case. Since we have three cases per each data point, three predictions are generated. The final prediction of the model is determined by a voting mechanism, i.e, class with highest number of votes will be selected as the final class label predicted for the input sample.

4 Experiments

In this section, we conduct comprehensive experiments to demonstrate the effectiveness of our proposed Causal-MMC in providing more accurate classification results for disaster relief. We first explain our experimental settings (Subsection 4.1). We then discuss our experimental results in subsection 4.2.

4.1 Experimental Setting

Dataset We evaluate our framework on the CrisisMMD dataset [3], which to the best of our knowledge is the only multi-modal dataset for disaster relief. CrisisMMD is a human-labeled multimodal dataset which consists of pairs of image-texts collected from Twitter during seven natural disasters in 2017. The dataset consists of three tasks: (1)Informativeness: which aims to determine if an image-text pair provides useful information for humanitarian aid during emergencies; (2)Humanitarian Event Classification: which aims to identify the category of the emergency events for an image-text pair; and (3)Severity of damages:

which determines the severity of the damage assessed from the image and the text. Since the images and tweets are labeled independently, some co-occurring tweets and images can have different labels for the same task. To overcome this issue, following existing works [19,25], we use a subset of the original data in which texts and images have same labels for a given task. Also, we utilize the same train-validation-test split ratio (70:15:15) as previous works [19,25].

Baselines We compare our results with three baselines: (1)Text-only: a uni-modal framework which is trained only on texts. The model uses the same architecture as our text embedding extractor; (2) Image-only: a uni-modal framework which is trained on the images of the Crisis-MMD; and (3) Concatenation-based Multi-modal classification (Concat-MMC) [19]: this frameworks utilizes both text and image modalities and integrates them by concatenating the extracted features from text and image. It has the same image feature and text embedding extractors as Causal-MMC and only differs in the data fusion mechanism from our work (uses concatenation instead of causal data fusion mechanism).

Evaluation Metrics To evaluate the performance of Causal-MMC and the baselines, we utilize well-known metrics used in previous studies [19,25]. In particular, we calculate and report Accuracy, Precision, Recall, and F1-score. Note that due to imbalanced nature of the CrisisMMD dataset for different classes, F1-score along with accuracy provide proper evaluation of our framework.

Table 1: Evaluation on Informativeness, Humanitarian, and Severity tasks.

Training Data	Model	Informativeness Task				Humanitarian Task				Severity Task			
		Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
Uni-Modal	Text-only	80.8	81.0	81.0	80.9	70.4	70.0	70.0	67.7	52.2	52.0	51.8	51.9
	Image-only	83.3	83.1	83.3	83.2	76.8	76.4	76.8	76.3	64.0	64.1	64.2	64.1
Multi-Modal	Concat-MMC[19]	84.4	84.1	84.0	84.2	78.4	78.5	78.0	78.3	67.0	68.0	67.0	59.0
	Causal-MMC(Our Model)	85.5	85.5	86.0	85.1	81.9	82.1	82.0	81.6	84.9	87.8	85.0	83.8

4.2 Experimental Results

Table 1 shows the experimental results for our proposed Causal-MMC and the baselines on Informativeness, humanitarian, and severity tasks. As can be seen from the table, Causal-MMC outperforms all baselines including uni-modal and multi-modal methods in terms of all evaluation metrics. We elaborate our in-depth observations as follows: (1) Multi-modal models (i.e., Concat-MMC and Causal-MMC), which use both text and image modalities outperform uni-modal models which only use either text or image data;(2) Causal-MMC outperforms the multi-modal concatenation-based baseline (Concat-MMC) which demonstrates the effectiveness of our proposed causal data fusion mechanism in achieving better classification performance; (3) Among uni-modal models, image-only framework demonstrates the best performance, which can be because the image feature extractor is more complex compared to the text embedding extractor

and thus it is able to capture more complicated patterns and demonstrate better performance. We believe replacing the text embedding extractor with more complicated state-of-the-arts can improve its performance.

To better analyze the behavior of our proposed causal data fusion mechanism, we further break down the performance of Causal-MMC based on three causal data fusion scenarios and report the results in Table 2. Our results show that in most cases, the performance of three causal graphs are mostly similar indicating that all three causal graphs are equally likely to generate the observed data. For the severity task, which is shown to be a more difficult task in terms of classification, the first causal graph shows considerably better performance, which indicates that the causal graph presented for case 1 (Figure 3a) is highly likely to be the causal graph generating the data for the third task, therefore most modalities for this task provide complementary information. This also highlights the importance of modeling the relations between the data modalities and considering all types of relations while performing the classification.

Table 2: Break down of causal data fusion cases for Informativeness, Humanitarian, and Severity tasks.

Model	Informativeness Task				Humanitarian Task				Severity Task			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
Case 1	85.3	85.0	85.0	85.0	78.3	79.7	78.0	78.4	99.8	99.8	100.0	99.8
Case 2	83.3	83.34	83.0	83.3	79.9	80.15	80.0	79.9	67.9	69.4	68.0	59.3
Case 3	84.0	83.9	84.0	84.0	78.8	78.5	79.0	78.4	67.4	68.8	67.0	59.1
Causal-MMC (Ensemble)	85.5	85.5	86.0	85.1	81.9	82.1	82.0	81.6	84.9	87.8	85.0	83.8

5 Conclusion

In this paper we propose Causal-MMC, a multi-modal framework for disaster classification. Causal-MMC is equipped with a novel causal data fusion module which considers the relationships between modalities while integrating them, creates the most informative set of features which do not contain redundant information, and therefore improves the classification performance. We verify the effectiveness of our proposed method on Crisis-MMD, a widely-used multi-modal disaster-related dataset. Our experimental results show that Causal-MMC outperforms the uni-modal and multi-modal classification baselines previously developed for disaster relief. While our proposed method adopts simple feature extractors to demonstrate our proposed causal data fusion mechanism, the performance of our framework can be further improved by utilizing state-of-the-art language models. We aim to explore this in our future research.

6 Acknowledgment

This work was supported by ARL under Award No. W911NF2020124. Opinions, interpretations, conclusions, and recommendations are those of the authors.

References

1. Abavisani, M., Wu, L., Hu, S., Tetreault, J., Jaimes, A.: Multimodal categorization of crisis events in social media. In: Proceedings of CVPR (2020)
2. Alam, F., Joty, S., Imran, M.: Graph based semi-supervised learning with convolution neural networks to classify crisis related tweets. In: AAAI (2018)
3. Alam, F., Ofli, F., Imran, M.: Crisismmd: Multimodal twitter datasets from natural disasters. In: Proceeding of ICWSM (2018)
4. Alam, F., Ofli, F., Imran, M.: Processing social media images by combining human and machine computing during crises. INT J HUM-COMPUT INT (2018)
5. Bai, H., Yu, G.: A weibo-based approach to disaster informatics: monitor in post-disaster situation via weibo text negative sentiment analysis. Nat Hazards (2016)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. ACL (2019)
8. Ding, C., Peng, H.: Minimum redundancy feature selection from microarray gene expression data. Journal of bioinformatics and computational biology (2005)
9. Gautam, A.K., Misra, L., Kumar, A., Misra, K., Aggarwal, S., Shah, R.R.: Multimodal analysis of disaster tweets. In: IEEE BigMM. IEEE (2019)
10. Imran, M., Castillo, C., Diaz, F., Vieweg, S.: Processing social media messages in mass emergency: A survey. ACM Computing Surveys (CSUR) (2015)
11. Körding, K., Tenenbaum, J.: Causal inference in sensorimotor integration. NIPs
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. NIPs (2012)
13. Kumar, S., Barbier, G., Abbasi, M., Liu, H.: Tweettracker: An analysis tool for humanitarian and disaster relief. In: Proceedings of ICWSM (2011)
14. Li, X., Caragea, D., Zhang, H., Imran, M.: Localizing and quantifying damage in social media images. In: ASONAM. IEEE (2018)
15. Nabi Abdolyousefi, R., et al.: Causal Inference Methods For Bias Correction In Data Analyses. Ph.D. thesis, Johns Hopkins University (2021)
16. Nguyen, D.T., Ofli, F., Imran, M., Mitra, P.: Damage assessment from social media imagery data during disasters. In: Proceedings of ASONAM (2017)
17. Nguyen, D.T., Al Mannai, K.A., Joty, S., Sajjad, H., Imran, M., Mitra, P.: Robust classification of crisis-related data on social networks using convolutional neural networks. In: ICWSM (2017)
18. Nguyen, D.T.e.a.: Automatic image filtering on social networks using deep learning and perceptual hashing during crises. arXiv preprint arXiv:1704.02602 (2017)
19. Ofli, F., Alam, F., Imran, M.: Analysis of social media data using multimodal deep learning for disaster response. arXiv preprint arXiv:2004.11838 (2020)
20. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
21. Sosea, T., Sirbu, I., Caragea, C., Caragea, D., Rebedea, T.: Using the image-text relationship to improve multimodal disaster tweet classification. In: ISCRAM(2021)
22. Sun, L.e.a.: Riva: a pre-trained tweet multimodal model based on text-image relation for multimodal ner. In: Proceedings of COLING (2020)
23. Weiss, K., Khoshgoftaar, T.M., Wang, D.: A survey of transfer learning. Journal of Big data **3**(1), 1–40 (2016)
24. Wu, Y., Chang, E.Y., Tseng, B.L.: Multimodal metadata fusion using causal strength. In: ACM SIGMM (2005)
25. Zou, Z., Gan, H., Huang, Q., Cai, T., Cao, K.: Disaster image classification by fusing multimodal social media data. ISPRS (2021)