GSMDA: GCN-Based Semi-Supervised Multi-Modal Domain Adaptation for Real-Time Disaster Detection

Yingdong Gou (✉ Gouyingdong@163.com)  
Northwest Institute of Mechanical and Electrical Engineering

Kexin Wang (✉ kexinwang0618@163.com)  
Northwest Institute of Mechanical and Electrical Engineering

Siwen Wei (✉ siwen_wei@stu.xidian.edu.cn)  
Xidian University

Changxin Shi (✉ shihchanghsin@163.com)  
Xidian University

Research Article

Keywords:

DOI: https://doi.org/

License: © This work is licensed under a Creative Commons Attribution 4.0 International License.  
Read Full License

Additional Declarations: No competing interests reported.
GSMDA: GCN-Based Semi-Supervised Multi-Modal Domain Adaptation for Real-Time Disaster Detection

Yingdong Gou¹, Kexin Wang¹*, Siwen Wei² and Changxin Shi²

¹Northwest Institute of Mechanical and Electrical Engineering, Xianyang, 712000, China.
²School of Computer Science and Technology, Xidian University, Xi’an, 710071, China.

*Corresponding author(s). E-mail(s): kexinwang0618@163.com; Gouyingdong@163.com; siwen_wei@stu.xidian.edu.cn; shihchanghsin@163.com;

Abstract

Nowadays, with the rapid expansion of social media as a means of quick communication, real-time disaster information is widely disseminated through these platforms. Determining which real-time and multi-modal disaster information can effectively support humanitarian aid has become a major challenge. In this paper, we propose novel end-to-end model, named GCN-based Semi-supervised Multi-modal Domain Adaptation (GSMDA), which consists of three essential modules: the GCN-based feature extraction module, the attention-based fusion module, and the MMD domain adaptation module. The GCN-based feature extraction module integrates text and image representations through GCNs, while the attention-based fusion module then merges these multi-modal representations using an attention mechanism. Finally, the MMD domain adaptation module is utilized to alleviate the dependence of GSMDA on source domain events by computing the maximum mean discrepancy across domains. Our experiments results demonstrate that GSMDA outperforms the current state-of-the-art models in terms of performance and stability.

Keywords: Graph convolutional network, Multi-modal disasters detection, domain adaptation
1 Introduction

The real-time disaster detection task plays a crucial role in extracting information as its aim is to identify useful multi-modal disaster information from social media, which can support humanitarian aid efforts. Timely analysis of the disaster’s location and infrastructure damage through this information can minimize casualties and economic losses [1]. As shown in Figure 1, the real-time disaster detection system requires training on previous events in the source domain and application to newly acquired multi-modal real-time disaster information in the target domain. The purpose of this task is to categorize the information into either ”Informative” or ”Not-Informative.”

Among existing methods for real-time disaster event detection, there has been a tendency to rely on only a single modality of information, disregarding the potential benefits of incorporating multiple modalities. For example, previous research [2] only utilized text information, ignoring images commonly shared on social media platforms. Conversely, other studies [3] placed emphasis on the significance of image information, utilizing it to evaluate the extent of environmental and infrastructure damage caused by disaster events. Nevertheless, it has been demonstrated that combining multiple sources leads to improved disaster event detection performance. Research [4] incorporated both text and image information in disaster events and implemented a convolutional neural network-based shared representation multi-modal deep learning architecture. Results of the study indicate that multi-modal training models outperformed single-modality models in disaster detection.

However, the current disaster event detection approach based on different sources is dependent on the accuracy of labeled data in the disaster event dataset, and does not account for unpredictable, unlabeled real-time disaster events on social media. [5] extract different modal features through the

Fig. 1  Tweet text and image pairs from different disastrous events
pre-training model, and employ the MME training strategy to achieve multi-modal domain adaptation, without solving the differences between different modalities.

To overcome this problem, this paper proposes a GCN-based Semi-supervised Multi-modal Domain Adaptation for real-time disaster detection, called GSMDA. GSMDA utilizes Graph Convolutional Network (GCN) to extract semantic features from different modalities and employs an attention mechanism to effectively integrate the multi-modal features. Furthermore, the method calculates the maximum mean discrepancy between the multi-modal features in different domains, reducing the dependence on the source domain events. Through comparative experiments with state-of-the-art baselines, the superiority of the proposed method has been demonstrated.

The main contributions of our paper are as follows:
1. We propose a novel end-to-end model, GCN-Based Semi-Supervised Multi-Modal Domain Adaptation (GSMDA) for real-time disaster detection, which harnesses the potential of unlabeled multi-modal data in real-time disaster events.
2. We utilize an attention mechanism to integrate the multi-modal features extracted using GCN representation, and implement domain adaptation through Maximum Mean Discrepancy (MMD).
3. We evaluate our model on three diverse disaster event datasets and demonstrate its superior performance compared to state-of-the-art multi-modal models.

2 Related work

2.1 Disaster Events Detection

The traditional field of disaster event detection has mainly focused on analyzing text information on social media, recognizing its significance for disaster response efforts [6]. Nguyen et al. applied a Text-CNN model to effectively classify disaster events on social media [3]. Similarly, Rosenthal et al. proposed a method to detect disaster events based on emotional analysis using Long-Short-Term Memory Networks (LSTM) [7]. With the development of convolutional neural networks (CNN), disaster event detection research has shifted towards analyzing image information on social media. For example, [8] utilized deep learning and perceptual hashing to detect disaster events on social media. [9] proposed a CNN-based image processing pipeline to extract meaningful information from disaster-related images shared on social media and classify them based on the level of damage.

In the multi-modal analysis field, there have been attempts to leverage both text and image information for disaster event detection. Mouzannar et al. [10] proposed a multi-modal deep learning framework that combined Text-CNN and Inception v4 to extract text and image features, respectively, and employs decision fusion to identify information related to damage, such as building
damage and landslides. Ferda et al. [4] developed a shared representation multi-modal deep learning architecture using Text-CNN and VGG16 and found that the multi-modal approach outperforms single-modal methods.

In the real-time disaster event detection, there is a recent interest in utilizing multiple modalities from social media to detect such events. To take full advantage of real-time unlabeled multi-modal data, [5] proposed a multi-modal feature representation based on BERT and VGG19 and implemented multi-modal domain adaptation using the multi-modal minimax entropy method. However, this method may not significantly improve the model performance when there is a significant difference between the feature distributions of the source and target domains. In contrast, the proposed GCN representation and attention fusion module in this paper overcomes the partial differences between different modalities, as well as by calculating the maximum mean diversity of multi-modal features across domains to reduce dependence on the source domain events.

2.2 Graph Convolution Networks

Graph Convolutional Network (GCN) is a powerful tool in the field of graph neural networks [11]. In recent years, GCN has been successfully applied to various natural language problems, such as text classification [12] and relationship extraction [13]. The early GCN study [14] introduced the concept of structured graph to obtain the neighbor state of nodes, and then to obtain the structured representation of all nodes. [15] proposed a dual gated graph attention network with dynamic iterative training to increase the depth of GCN layers. [16] utilized different types of attention maps and dependency labels to learn multi-level syntactic representations.

The GCN method employed in this paper is inspired by [17], which employs multi-head attention calculation based on BERT to enhance the performance of GCN. Additionally, to reduce the distribution difference between text modal features and image modal features, the attention map is constructed following the method of [18] to perform image modal feature extraction.

3 Method

The GSMDA model consists of three main components: the GCN-based feature extraction module, the attention-based feature fusion module, and the MME module, as depicted in Figure 2. Specifically, the input to the GSMDA model comes from both the source domain and target domain, where the target domain has a limited number of labeled data and a large amount of unlabeled data. The GCN-based feature extraction module is used to extract the multi-modal features from both text and image. Then, these multi-modal features are combined into a unified representation through the attention-based feature fusion module. Finally, the MMD domain adaptation module is utilized to alleviate the dependence of GSMDA on source domain events by computing the maximum mean discrepancy across domains.
3.1 GCN-based feature extraction module

3.1.1 Syntactic feature extraction based on Multi-head GCNs

Formally, the text content can be represented as a sequence of \( n \) words, denoted as \( W = [w_1, w_2, ..., w_n] \), where each word, \( w_i \), constitutes the \( i \)-th element of the sequence. To begin, we use a sentence encoder to convert the word sequence \( W \) into a hidden embedding representation:

\[
\{h_1, h_2, ..., h_n\} = T\{w_1, w_2, ..., w_n\} \tag{1}
\]

In the GSMDA model, \( T \) refers to the sentence encoder which utilizes the BERT [19]. Specifically, the input sequence for BERT is constructed as follows:

\[
S = [<CLS>, \text{sentence}, <SEP>] \tag{2}
\]

Since our text input consists of a single sentence, both the \(<CLS>\) and \(<SEP>\) are set to zero. After inputting the aforementioned input representation into 12 transformer, we obtain the context representation \( H_{\text{bert}} \) of BERT:

\[
H_{\text{bert}} = \{h_1, h_2, ..., h_n, h_{n+1}\} \tag{3}
\]
We obtain the final sentence representation $H_{\text{sentence}}$ by removing the representations of $<\text{CLS}>$ and $<\text{SEP}>$, which will serve as the nodes input to the subsequent Multi-head GCNs.

$$H_{\text{sentence}} = \{h_1, h_2, ..., h_n\}$$ (4)

To effectively encode both the text and image into a unified feature space, we adopt the technique of abstract meaning representation (AMR) [20] to extract crucial syntax information from the text. Specifically, AMR provides a comprehensive representation of the abstract meaning of text, with a granular representation of 150 semantic roles. We utilize the CAMR [21] parser, which is trained on the Stanford CoreNLP [22] dataset, to perform syntactic encoding.

The graph convolution network (GCN) framework can be described as follows: Given a sentence $W$ with $n$ nodes, there exists a syntax dependency tree $G = \{W, E\}$, where $E$ represents the set of relationships between nodes. For example, an edge $(w_i, w_j)$ represents a relationship where the word $w_i$ points to the word $w_j$ with a dependency label $K(w_i, w_j)$. For instance, in the Figure 3, there is an edge with the “nmod” type from the node “deploy” to the node “region”. The syntactic structure $E$ corresponding to the sentence is simplified to an adjacency matrix $A$. The size of adjacency matrix $A$ is $n \times n$. If there is an edge between nodes $w_i$ and $w_j$, then $a_{ij} = 1$, $a_{ji} = 1$, otherwise $a_{ij} = 0$, $a_{ji} = 0$, where $a_{ij}$ is an element in $A$.

The traditional GCN method utilizes the adjacency matrix $A$ and the original node’s BERT representation to generate multiple context-aware graph structures. Specifically, we gather information from the neighboring nodes of the adjacency matrix $A$ through the graph convolution operations. The $l_{\text{text}}$-th layer graph convolution vector, $h_i^{l_{\text{text}}}$, of node $i$ can be calculated as follows:

$$h_i^{l_{\text{text}}} = \rho \left( \sum_{j=1}^{n} A_{(i,j)} W^{(l_{\text{text}})} h_j^{(l_{\text{text}}-1)} + b^{(l_{\text{text}})} \right)$$ (5)

where, $W^{(l_{\text{text}})}$ and $b^{(l_{\text{text}})}$ represent the learnable weight matrix and the bias term, respectively, and $\rho$ represents a RELU activation function. In addition, we utilize the output of the sentence encoding module $H_{\text{sentence}}$ to initialize the node representation $h_{(i,j)}^{(0)}$ of the first layer GCNs.
However, this paper aims to leverage the BERT representation more effectively and thus employs attention mechanism \cite{23} to calculate the attention adjacency matrix $A^*$ instead of $A$. Specifically, we leverage the BERT semantic vector to determine the correlation between words from various perspectives (N-head):

$$A^*_{(l_{\text{text}}, n)} = \text{softmax} \left( \frac{Q^{l_{\text{text}}-1} W_n^{l_{\text{text}}-1} * \left( K^{l_{\text{text}}-1} W_n^{l_{\text{text}}-1} \right)}{\sqrt{d}} \right) A \quad (6)$$

where both $Q^{l_{\text{text}}-1}$ and $K^{l_{\text{text}}-1}$ denote the outputs of syntactic graph representation $H$. $W_n^{l_{\text{text}}-1}$ and $W_n^{l_{\text{text}}-1}$ denote parameter matrices and $d$ denotes the dimension of the hidden layer.

Therefore, for each layer of the Multi-head GCNs, we can obtain $N$ distinct types of attention node information via $A^*_{(l_{\text{text}}, n)}$. Subsequently, we fusion these different representations of attention nodes in $l_{\text{text}}$-th layer of the Multi-head GCNs to produce the output $out_{\text{text}}^{l_{\text{text}}}_{i}$:

$$out_{\text{text}}^{l_{\text{text}}}_{i} = W_{l_{\text{text}}}^{\text{tran}} (\text{contact}(h_1^{l_{\text{text}}}, h_2^{l_{\text{text}}}, ..., h_N^{l_{\text{text}}})) + b_{l_{\text{text}}}^{\text{tran}} \quad (7)$$

where $W_{l_{\text{text}}}^{\text{tran}}$ and $b_{l_{\text{text}}}^{\text{tran}}$ represent the learnable transformation weight matrix and the bias term, respectively, and $\text{contact}(.)$ represents the splicing operation.

Finally, after the $L_{\text{text}}$ layers of Multi-head GCNs, we can obtain the text syntax features $H^{L_{\text{text}}}$ as follows:

$$H^{L_{\text{text}}} = [out_{\text{text}}^{1}_{i}, out_{\text{text}}^{2}_{i}, ..., out_{\text{text}}^{L_{\text{text}}}_{i}] \quad (8)$$

### 3.1.2 Attention-based GCNs

In order to obtain a image graph that resembles text from an image, we first utilize the state-of-the-art object detection model \cite{24} for identifying image sub-nodes (sub-object). Subsequently, we employ the pre-trained VGG19 to extract features from each sub-node and the entire image, which are represented as $g_{\text{node}}$ and $G$, respectively. For each sub-node feature $g_i$ embedded in $E$, we formulate an attention query vector $Q_{\text{node}}$ by integrating it with the contextual information from the image features.

$$Q_{\text{node}} = W_Q [g_{\text{node}}; E] + b_Q \quad (9)$$

where, $W_Q$ and $b_Q$ represent the learnable transformation weight matrix and the partial term, respectively.

Finally, we utilize the pre-trained VGG19 to extract multiple convolutional feature maps of 14x14 regions in the image, which serve as the attention key values for local regions, denoted as $K$. Meanwhile, we construct an image
attention map corresponding to the text syntax graph, which is denoted as $A^\text{img}_{(i,j)}$.

$$A^\text{img}_{(i,j)} = \text{softmax}(K_i \ast Q_{\text{node}}) = \frac{\exp(K_i \ast Q_{\text{node}})}{\sum_{j \in 14 \times 14} \exp(K_j \ast Q_{\text{node}})} \quad (10)$$

It is worth mentioning that we employ the method of [18] to pre-train the images in the IMSTU dataset to generate a more accurate image attention graph. Similarly, for the $l_{\text{img}}$-th layer of GCN, we obtain the attention graph feature of the image, denoted as $g^l_{\text{node,i}}$.

$$g^l_{\text{node,i}} = \rho\left( \sum_{\text{node,j}} A^\text{img}_{(i,j)} W^{l_{\text{img}}} g^{l_{\text{img}}-1}_{\text{node,j}} + b^{l_{\text{img}}_0} \right) \quad (11)$$

where, $W^{l_{\text{img}}}$ and $b^{l_{\text{img}}_0}$ represent the learnable weight matrix and the bias term respectively, and $\rho$ represents a RELU activation function.

To enhance the performance of the Attention-based GCNs, we incorporate the residual mechanism as inspired by ResNet [25]. Specifically, we integrate the local feature representation of the image extracted by pre-trained VGG19 with the non-local contextual information of the image, extracted via the image attention graph.

$$g^l_{\text{node,i}} = \rho\left( \sum_{\text{node,j}} A^\text{img}_{(i,j)} W^{l_{\text{img}}} g^{l_{\text{img}}-1}_{\text{node,j}} + b^{l_{\text{img}}_0} \right) + E \quad (12)$$

Where $E$ represents the local feature of the entire image extracted using pre-trained VGG19, which is only integrated into the previous $L_{\text{img}}-1$ layers, and $\rho$ represents a RELU activation function.

Finally, after the $L_{\text{img}}$ layers of GCN, we can obtain the image attention graph features $G^{L_{\text{img}}}$ as follows:

$$G^{L_{\text{img}}} = [g^1_{\text{node,i}}, g^2_{\text{node,i}}, ..., g^{L_{\text{img}}}_{\text{node,i}}] \quad (13)$$

### 3.2 Attention-based fusion module

In order to reduce the discrepancy between text and image features, we adopt an attention mechanism and $l2$-norm to seamlessly integrate the text feature $H^{L_{\text{ext}}}$ with the image feature $G^{L_{\text{img}}}$, and the fusion feature is denoted as $F_{\text{att,l2}}$.

$$F_{\text{att,l2}} = l2(F_{\text{att}}) = \frac{F_{\text{att}}}{\max(\|F_{\text{att}}\|_2, \epsilon)} \quad (14)$$

$$F_{\text{att}} = \sum_{l=1}^{L_{\text{ext}}+L_{\text{img}}} v^l F^{\{\text{text,img}\}} \quad (15)$$

$$F^{\{\text{text,img}\}} = \text{contact}(H^{L_{\text{ext}}}, G^{L_{\text{img}}}) \quad (16)$$
\[ v^l_i = \text{softmax}(f^l_i) = \frac{e^{	ext{exp}(f^l_i)^T\text{ctx}}}{\sum_{j=1}^{L_{\text{text}}+L_{\text{img}}} e^{	ext{exp}(f^l_i)^T\text{ctx}}} \quad (17) \]

\[ f^l_i = \text{tanh}(W_{\text{att}}f^l_i + b_{\text{att}}) \quad (18) \]

Where \( v^l_i \) denotes the normalized attention weight represented by the fusion layer of the \( l \)-th level; \( f^l_i \) is the node representation of layer \( l \) of \( F^{\{\text{text}, \text{img}\}} \); \( W_{\text{att}} \) and \( b_{\text{att}} \) are the learnable weight matrices and bias terms, respectively; \( \text{ctx} \) is a randomly initialized context vector [26], and \( \text{contact}(\cdot) \) represents the splicing operation.

### 3.3 MMD domain adaptation module

Finally, to achieve multi-modal domain adaptation, the maximum mean discrepancy (MMD) [27] loss is employed to minimize the difference between the source domain \( (D_s) \) and target domain \( (D_t) \) features, improving the generalization performance of the model on the target domain. The MMD between \( D_s \) and \( D_t \) features can be calculated as follows:

\[
\text{MMD}(D_s, D_t) = \left\| \frac{1}{D_s} \sum_{F_{\text{att},l2} \in D_s} \varphi(F_{\text{att},l2}) - \frac{1}{D_t} \sum_{F_{\text{att},l2} \in D_t} \varphi(F_{\text{att},l2}) \right\| \quad (19)
\]

Where, \( \varphi(\cdot) \) represents the formula [28] for calculating the maximum average difference.

It is worth noting that our goal is to minimize the difference between domains and obtain a representation that is conducive to training strong classifiers. Therefore, we set the objective function as follows:

\[
\mathcal{L}_{\text{MMD}} = \mathcal{L}_{\text{cls}}(D_s, D_t, y) + \lambda \text{MMD}^2(D_s, D_t) \quad (20)
\]

\[
\mathcal{L}_{\text{cls}}(D_s, D_t, y) = \mathbb{E}_{(x,y) \in D_s, D_t} \mathcal{L}_{\text{ce}}(p(x), y) \quad (21)
\]

Where \( \mathcal{L}_{\text{ce}}(\cdot) \) represents the cross-entropy loss function, \( y \) is the label of multimodal data in the source domain \( D_s \) and target domain \( D_t \). \( \lambda \) is the regularization hyper-parameter. This approach enables our model to effectively leverage the abundant labeled data in the source domain to improve its performance on the target domain.

### 4 Experiment

#### 4.1 Dataset Descriptions

In this paper, the performance of GSMDA is evaluated using the CrisisMMD [29] multi-modal dataset. The CrisisMMD consists of multi-modal data collected from Twitter using keywords and topic tags of specific events, including
seven natural disasters that occurred in 2017. The details of the dataset can be found in Table 1.

<table>
<thead>
<tr>
<th>Disaster Event</th>
<th>Keywords/Hastag</th>
<th>Start Date</th>
<th>End Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hurricane Irma</td>
<td>Hurricane Irma, Irma storm,</td>
<td>Sep 06, 2017</td>
<td>Sep 21, 2017</td>
</tr>
<tr>
<td>Hurricane Harvey</td>
<td>Hurricane Harvey, Hurricane,</td>
<td>Aug 26, 2017</td>
<td>Sep 20, 2017</td>
</tr>
<tr>
<td>Hurricane Maria</td>
<td>Hurricane Maria, Maria Cyclone,</td>
<td>Sep 20, 2017</td>
<td>Nov 13, 2017</td>
</tr>
<tr>
<td>Mexico earthquake</td>
<td>Mexico earthquake, earthquake,</td>
<td>Sep 20, 2017</td>
<td>Oct 06, 2017</td>
</tr>
<tr>
<td>California wildfires</td>
<td>California fire, California wildfires,</td>
<td>Oct 10, 2017</td>
<td>Oct 27, 2017</td>
</tr>
<tr>
<td>Iraq earthquake</td>
<td>Iran/halabja earthquake,</td>
<td>Nov 13, 2017</td>
<td>Nov 19, 2017</td>
</tr>
</tbody>
</table>

Considering the need for multi-modal domain adaptation across different events, we align our data with previous research in this field. Specifically, we chose three events, Hurricane Harvey (HUH), California Wildfire (CWF), and Mexico Earthquake (MEQ), for testing purposes. Table 2 provides more details on the event.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Start Date</th>
<th>End Date</th>
<th>informative</th>
<th>Not-informative</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUH</td>
<td>Aug 26, 2017</td>
<td>Sep 20, 2017</td>
<td>2615</td>
<td>2836</td>
</tr>
<tr>
<td>CWF</td>
<td>Sep 20, 2017</td>
<td>Oct 6, 2017</td>
<td>1352</td>
<td>1361</td>
</tr>
</tbody>
</table>

The purpose of our model is to simultaneously train and perform multi-modal domain adaptation in the source and target domains, and evaluate its performance in the target domain. If the multi-modal input (text-image pair) from the target domain is effective for disaster response or humanitarian aid, it should be classified as "informative", otherwise as "not-informative".

4.2 Experiment Setting

For the text data in the CrisisMMD dataset, we follow the standard text pre-processing procedure as commonly used in related research [30]. For the image data, we resize the images to 224x224x3 and then perform the typical
preprocessing steps, which include scaling the image pixels between 0 and 1 and normalizing each channel based on the ImageNet dataset.

In the feature extraction module based on GCN, the text features are obtained from a pre-trained \( BERT_{Base} \) model, while the image features are extracted from a VGG-19 model that was pre-trained on the ImageNet dataset \([31]\). To avoid over-fitting, the parameters of both \( BERT_{Base} \) and VGG-19 are kept frozen during training. The model is trained for 100 epochs in 24 batches, with a learning rate of 0.001, and Adam \([32]\) optimizer is employed to optimize the parameters in BDANN. Finally, the regularization hyper-parameter is set to \( \lambda = 0.25 \).

### 4.3 Experimental results and analysis

In order to evaluate the classification performance of GSMDA on the CrisisMMD dataset, we have collected comparable models, including text-only, image-only, Att-RNN \([33]\) and BSSDA.

1. **Text-Only**
   
   Text-Only only uses the text as input. The syntactic features are first extracted from the pre-trained \( BERT_{Base} \) model and Multi-head GCNs, and then these features are directly classified into the disaster events.

2. **Image-Only**
   
   Image-Only only uses the image as input. The image attention graph features are first extracted from the pre-trained VGG-19 and Attention-based GCNs, then these features are directly classified into the disaster events.

3. **Att-RNN**
   
   Att-RNN leverages LSTM to extract tweet text features and pre-trained deep CNN to extract image features, in order to obtain a joint representation for disaster event detection.

4. **BSSDA**
   
   BSSDA extracts text features from a pre-trained \( BERT_{Base} \) model and image features from a pre-trained VGG-19 model to obtain a joint representation, and employs the MME method \([34]\) to achieve multi-modal domain adaptation for disaster event detection.

5. **GSMDA-att**
   
   GSMDA-att directly classifies disaster events by concatenating both text and image features, without using an attention-based fusion strategy.

   Table 3 and Table 4 present the experimental results of our GSMDA model. The performance evaluation indicators used to assess the model’s performance include precision (P), recall(R), and F1 score (F1). The first row of Table 3
displays our domain adaptation method in the “Method” column. The “Transfer” block represents that the model was trained directly in the source domain and evaluated on the target domain, while the “MMD” block indicates that the model was trained using our domain adaptation method. The experimental results in the CWF to HUH column indicate that the source domain is a CWF event and the target domain is a HUH event.

Table 3 Performance of BSSDA against baselines on multimodal datasets.

<table>
<thead>
<tr>
<th>Net</th>
<th>Method</th>
<th>HUH to CWF</th>
<th></th>
<th></th>
<th>CWF to HUH</th>
<th></th>
<th></th>
<th>HUH to MEQ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td>P</td>
</tr>
<tr>
<td>TextOnly</td>
<td></td>
<td>64.7</td>
<td>60.6</td>
<td>62.6</td>
<td>64.2</td>
<td>59.8</td>
<td>61.9</td>
<td>62.6</td>
</tr>
<tr>
<td>Image-Only</td>
<td>Transfer</td>
<td>63.8</td>
<td>59.7</td>
<td>61.7</td>
<td>61.7</td>
<td>58.3</td>
<td>59.9</td>
<td>59.3</td>
</tr>
<tr>
<td>Att-RNN</td>
<td></td>
<td>65.2</td>
<td>60.5</td>
<td>62.7</td>
<td>63.4</td>
<td>60.2</td>
<td>61.8</td>
<td>62.1</td>
</tr>
<tr>
<td>TextOnly</td>
<td></td>
<td>70.8</td>
<td>68.3</td>
<td>69.</td>
<td>71.4</td>
<td>68.3</td>
<td>69.8</td>
<td>70.7</td>
</tr>
<tr>
<td>Image-Only</td>
<td>MMD</td>
<td>69.7</td>
<td>68.2</td>
<td>68.9</td>
<td>70.8</td>
<td>68.1</td>
<td>69.4</td>
<td>69.8</td>
</tr>
<tr>
<td>Att-RNN</td>
<td></td>
<td>70.1</td>
<td>69.1</td>
<td>69.6</td>
<td>69.9</td>
<td>68.2</td>
<td>69.0</td>
<td>68.3</td>
</tr>
<tr>
<td>BSSDA</td>
<td></td>
<td>73.8</td>
<td>69.9</td>
<td>71.8</td>
<td>72.4</td>
<td>69.7</td>
<td>71.0</td>
<td>71.1</td>
</tr>
<tr>
<td>GMSDA-att</td>
<td></td>
<td>74.2</td>
<td>72.9</td>
<td>73.5</td>
<td>73.9</td>
<td>72.6</td>
<td>73.2</td>
<td>74.1</td>
</tr>
<tr>
<td>GMSDA</td>
<td></td>
<td>74.7</td>
<td>73.5</td>
<td>74.1</td>
<td>75.2</td>
<td>72.7</td>
<td><strong>73.9</strong></td>
<td>74.8</td>
</tr>
</tbody>
</table>

As evidenced by the results in Table 3 and Table 4, the model that utilized the MMD method has achieved superior performance in terms of precision, recall, and F1, demonstrating the effectiveness of the MMD training strategy. However, while TextOnly and ImageOnly combined with MMD achieved good results, they still fall short compared to the multi-modal method BSSDA, which leverages both text and visual features. On the other hand, The F1 score of GSMDA outperforms BSSDA by approximately 3% in all tasks, highlighting the effectiveness of our GCN-based unified representation and MMD-based domain adaptation strategy. Furthermore, the significant increase in the F1 of GSMDA compared to GSMDA-att indicates that proper fusion strategies play a crucial role in enhancing the performance of multi-modal domain adaptation task.

It is worth noting that the results of a single experiment may not accurately reflect the true performance and generalizability of a model. In our six experimental groups, we found that the mean F1 of the GSMDA and BSSDA models are 74.0(variance is 0.218 ) and 71.2(variance is 0.298 ) , respectively. These results demonstrate the strong robustness of our GSMDA model in adapting to various multi-modal disaster events.
Table 4 Performance of BSSDA against baselines on multimodal datasets.

<table>
<thead>
<tr>
<th>Net</th>
<th>Method</th>
<th>CWF to MEQ</th>
<th>MEQ to CWF</th>
<th>MEQ to HUH</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
</tr>
<tr>
<td>TextOnly</td>
<td></td>
<td>61.9</td>
<td>59.2</td>
<td>60.5</td>
</tr>
<tr>
<td>Image-Only</td>
<td>Transfer</td>
<td>60.9</td>
<td>59.8</td>
<td>60.3</td>
</tr>
<tr>
<td>Att-RNN</td>
<td></td>
<td>62.1</td>
<td>59.4</td>
<td>60.7</td>
</tr>
<tr>
<td>TextOnly</td>
<td></td>
<td>69.9</td>
<td>67.1</td>
<td>68.5</td>
</tr>
<tr>
<td>Image-Only</td>
<td>MMD</td>
<td>67.9</td>
<td>66.0</td>
<td>66.9</td>
</tr>
<tr>
<td>Att-RNN</td>
<td></td>
<td>69.1</td>
<td>67.1</td>
<td>68.1</td>
</tr>
<tr>
<td>BSSDA</td>
<td></td>
<td>72.1</td>
<td>70.0</td>
<td>71.0</td>
</tr>
<tr>
<td>GMSDA-att</td>
<td></td>
<td>74.8</td>
<td>71.7</td>
<td>73.2</td>
</tr>
<tr>
<td>GMSDA</td>
<td></td>
<td>75.4</td>
<td>72.3</td>
<td>73.8</td>
</tr>
</tbody>
</table>

4.4 Parameter Analysis

When implementing the MMD method to assess the similarity between the source and target domains, it is necessary to set the regularization hyperparameter $\lambda$ to balance the training loss. If $\lambda$ is set too low, the MMD strategy will have minimal influence on the model’s loss during iteration. Conversely, if $\lambda$ is set too high, the model will be over-regularized and learn a degenerate representation in which all points are too similar. To find the optimal results, we conducted experiments by varying $\lambda$ from 0.1 to 0.4, as presented in Table 5. Our results show that when the regularization parameter was set to $\lambda = 0.25$, the GMSDA model achieved the best performance, striking a balance between the model’s ability to make accurate predictions and avoiding over-fitting.

Table 5 Model performance on variants of $\lambda$.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\lambda$</th>
<th>HUH to CWF</th>
<th>CWF to HUH</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>72.1</td>
<td>71.8</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>73.8</td>
<td>72.9</td>
</tr>
<tr>
<td>BSSDA</td>
<td>0.25</td>
<td>74.7</td>
<td>73.5</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>74.4</td>
<td>73.0</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>73.3</td>
<td>72.7</td>
</tr>
</tbody>
</table>
5 Conclusion

In this paper, we proposed GMSDA, a GCN-based semi-supervised multi-modal domain adaptation for real-time disaster detection. GMSDA consists of three main modules, a GCN-based feature extraction module, a GCN-based feature extraction module, an attention-based fusion module, and a Maximum Mean Discrepancy (MMD) domain adaptation module. First, the GCN-based feature extraction extracts the multi-modal features from the multi-modal data. Then, the attention-based fusion module combines these multi-modal representations using an attention mechanism and fed into the MMD domain adaptation module to reduce the dependence of the GMSDA model on the source domain events. To evaluate the performance of the GMSDA model, we conduct experiments on three diverse disaster event datasets and demonstrate its superior performance compared to state-of-the-art multi-modal models. In the future, besides applying different strategies for feature fusion, we also plan to investigate incorporating additional modalities such as audio or video data to enhance the representation capability of GSMDA. Our ultimate goal is to develop a robust and scalable multi-modal disaster detection model that can accurately identify disaster events in real-time and provide critical information for disaster response and humanitarian aid.

References


GSMDA: GCN-Based Semi-Supervised Multi-Modal Domain Adaptation


Figures

Figure 1

Figure 2

Network structure of GCN-Based Semi-Supervised Multi-Modal Domain Adaptation (GSMDA)

Figure 3

\{\color{blue}\text{ Tweet text and image pairs from different disastrous events }\}

Figure 4

An example sentence with the CAMR dependency parsing result.

Figure 5

Figure 6