2022 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-8563-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICME52920.2022.9859834

HGLNET: A GENERIC HIERARCHICAL GLOBAL-LOCAL FEATURE FUSION NETWORK FOR MULTI-MODAL CLASSIFICATION

Jiesheng Wu, Junan Zhao, Jing Xu[*]

College of Artificial Intelligence, Nankai University, Tianjin 300350, China
{jasonwu,zja}@mail.nankai.edu.cn, xujing@nankai.edu.cn

ABSTRACT

Multi-modal fusion aims to capture the semantic interactions between different modalities for many downstream classification tasks. However, previous work usually considers that each modality contributes equal information to the final classification and extracts the global features of each modality for fusion. In this paper, inspired by these two observations, we propose a generic **H**ierarchical **G**lobal-**L**ocal feature fusion **N**etwork (HGLNet) for multi-modal classification. Specifically, HGLNet has three merits compared to the current work. (1) HGLNet proposes a **G**lobal **G**ated **A**ttention (GGA) module, which adaptively generates weights that represent the contributions of different modalities. (2) HGLNet presents a novel **C**ross **R**esidual **T**ransformer (CRT) module to capture the fine-grained local interactions. (3) HGLNet utilizes hierarchical information for multi-modal fusion. Extensive experiments on three public datasets demonstrate that HGLNet achieves competitive performance against the state-of-the-art methods for three kinds of multi-modal classification tasks.

Index Terms— Multi-modal fusion, global gated attention, cross residual transformer, classification

1. INTRODUCTION

With the development of social media (e.g., Weibo, Twitter), various large-scale multimedia data (e.g., text and images) have emerged on social media explosively. Meanwhile, this trend also brings unprecedented challenges for multi-modal classification [1, 2]. Currently, multi-modal classification tasks usually include multi-modal fake news detection [3], multi-modal sentiment analysis [4], multi-modal tweet classification [5, 6], etc., all of which are research hot-spots in the area of multimedia. For example, Figure 1 shows two tweets sampled from Weibo. We can observe that the prediction based on textual modality is correct in tweet (a). However, in tweet (b), the prediction based on visual modality is matched with the ground truth. Therefore, it cannot predict the correct label and fails to convey sufficient semantics for classification by only single-modality. In addition, we can notice that the probability value indicated by each modality for the corresponding label is unequal, which shows that the semantics contained in each modality are inconsistent. This paper aims at exploiting textual and visual information to tackle various downstream tasks of multi-modal classification.

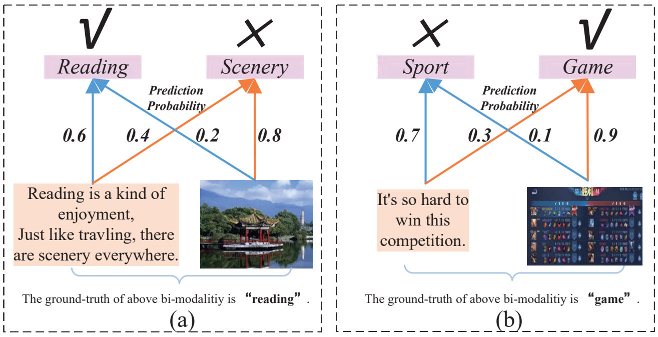
√
Reading
×
Scenery
Prediction Probability
0.6
0.4
0.2
0.8
Reading is a kind of enjoyment,
Just like travling, there are scenery everywhere.
The ground-truth of above bi-modalitiy is "reading".
(a)
×
Sport
√
Game
Prediction Probability
0.7
0.3
0.1
0.9
It's so hard to win this competition.
The ground-truth of above bi-modalitiy is "game".
(b)

Fig. 1. An example of multi-modal classification.

Compared to the traditional classification tasks, the main challenge of multi-modal classification tasks is how to capture multi-modal semantic interactions and learn discriminative fusion features. Since the rapid development of Deep Neural Networks (DNNs) [7], considerable researchers have recently employed powerful feature extraction techniques for multi-modal classification. For example, Jin et al. [8] use the Recurrent Neural Network (RNN) and attention mechanism [9] for multi-modal fake news detection. Truong et al. [10] propose a visual aspect attention network for multi-modal sentiment analysis.

Motivation. Although the previous methods have strikingly boosted the performance of multi-modal classification tasks, the fields remain with several unsolved challenges: (1) They usually treat textual and visual features equally from a global perspective and ignore the semantic knowledge contained in each modality is inconsistent, so that each modality contributes to the final classification differently; (2) most of them focus on capturing high-level local semantic interactions between textual and visual modality, failing to exploit the hierarchical semantic interactions efficiently to assist in learning a better fusion representation. To this end, we investigate how to represent the contribution value of each modality for a final decision and realize the hierarchical semantic interactions and fusions efficaciously to support multi-modal

[*]Corresponding author

classification tasks.

Our Method. To represent the contribution value of each modality for the final classification and capture the hierarchical local interactions between textual and visual modalities, we propose a generic **H**ierarchical **G**lobal-**L**ocal feature fusion **N**etwork (HGLNet). Specifically, HGLNet includes two core modules: the **G**lobal **G**lobal **A**ttention (GGA) fusion module and the **C**ross **R**esidual **T**ransformer (CRT) module. GGA generates two weights to represent the contribution values by fusing the global features of two modalities. Moreover, some previous work [3, 11–13] usually uses standard Transformer Encoder to model multi-modal fusion, which easily leads to too many parameters. In addition, in [11], they confirm that keeping the unique properties for each modality is essential while fusing multi-modal features. Thus, CRT proposes a novel **D**ual-stream **C**ross **R**esidual **S**elf-**A**ttention (DCRSA) mechanism to replace the traditional **S**elf-**A**ttention (SA) and capture the local semantic interactions better by fusing the local features of two modalities while maintaining the unique information of related modalities. Furthermore, to capture the hierarchical semantic interactions, HGLNet applies multiple CRT modules to fuse hierarchical context semantic interactions. To this end, HGLNet final concatenates all obtained fusion features for classification.

Contributions. The main contributions are summarized as follows:

- We propose a novel generic network for multi-modal classification tasks, named HGLNet, which integrates the hierarchical global and local information of each modality for classification.
- To the best of our knowledge, we are the first to exploit a GGA module to generate two weights for representing the contribution values of textual and visual modalities for the final classification. We further present a CRT module to capture the fine-grained local semantic interactions between two modalities better.
- Experimental results on datasets of three different tasks demonstrate that our method achieves competitive performance compared with state-of-the-art methods.

2. RELATED WORK

In this paper, we employ the proposed HGLNet in three multi-modal classification tasks: (1) multi-modal fake news detection, (2) multi-modal sentiment analysis, (3) multi-modal tweet classification. Thus, We will review the related works from the three aspects.

Multi-modal fake news detection. Multi-modal fake news detection can be defined as a binary classification task, which focuses on a tweet on social media is fake news or not [14]. Previous work mainly concentrates on single-modal fake news detection (e.g., text or images) [15, 16]. However, since the advance of multi-modal technologies, the performance of multi-modal fake news detection has been enhanced [8]. Song et al. [11] introduce a crossmodal attention residual and multichannel network for fake news detection. Qian et al. [3] propose a hierarchical multi-modal contextual attention network for detection. **Multi-modal sentiment analysis.** Recently, sentiment analysis has gradually developed in the multi-modal field. Sentiment analysis is usually regarded as a multi-class classification task. Truong et al. [10] present a visual aspect attention network for multi-modal review sentiment analysis. Xu et al. [17] propose a co-memory network for multi-modal sentiment analysis. **Multi-modal tweet classification.** Social media consists of text, images, and video. Therefore, it is essential for social media to classify tweets into correct labels. Recently, Abavisani et al. [5] present a novel multi-modal framework for classifying multi-modal data in the crisis domain. Hu et al. [6] are the first to introduce a novel social network multi-modal classification dataset, which includes 18 general categories.

3. METHODOLOGY

Multi-modal classification tasks are defined as a binary-class or multi-class classification problems, focusing on whether samples on multi-modal data are incorrect labels. In this section, a generic **H**ierarchical **G**lobal-**L**ocal feature fusion **N**etwork (HGLNet) is proposed. Figure 2 shows an overview of HGLNet. HGLNet consists of three stages: (1) *multi-modal feature embedding*, (2) *multi-modal feature fusion*, and (3) *classification*.

3.1. Multi-modal Feature Embedding

Given a multi-modal sample S from multimedia consists of text and attached images, the model will output $Y = \{y_1, \ldots, y_i\}$ to indicate to the label of the sample, where $Y = y_i$ denotes that the piece belongs to the y_i category. Just like this, the input of our model is a multi-modal sample $S = \{T, V\}$, where T and V denote the textual content and visual content, respectively.

Text Encoder. To better model the semantic information of text, we consider the global and local information of text: textual content embedding and word embeddings. We employ the pre-trained BERT [18] to embed the textual content. First, given a text T, we model T as a word sequences $T = \{w_1, \ldots, w_m\}$ (m represents the number of words), we further denote the word embeddings and textual content embedding as follows:

$$\mathbf{E} = \mathrm{BERT}(T) = \{\mathbf{e}_{\mathrm{cls}}, \mathbf{e}_1, \ldots, \mathbf{e}_m\}, \tag{1}$$

where $\mathbf{e}_m \in \mathbb{R}^{d_t}$ denotes the hidden embedding of the corresponding token in BERT, and d_t denotes the dimension of the word embeddings. In Equation 1, $\mathbf{e}_{\mathrm{cls}}$ is the [CLS] token embedding in BERT that represents the global information of the textual content.

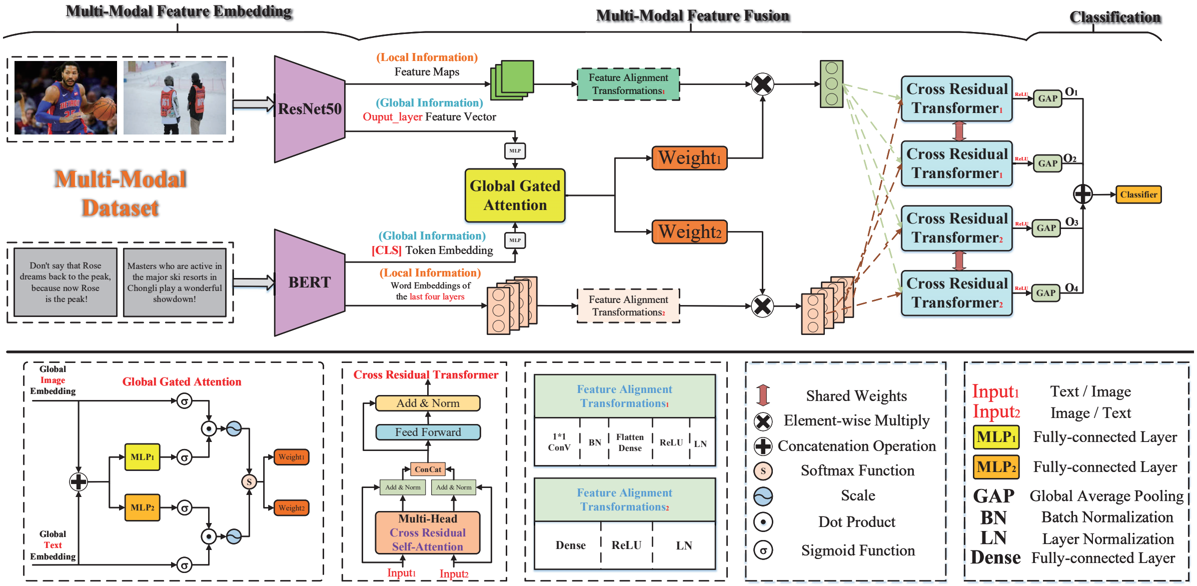
Multi-Modal Feature Embedding
Multi-Modal Feature Fusion
Classification
Multi-Modal Dataset
ResNet50
BERT
Don't say that Rose dreams back to the peak, because now Rose is the peak!
Masters who are active in the major ski resorts in Chongli play a wonderful showdown!
(Local Information)
Feature Maps
(Global Information)
Ouput_layer Feature Vector
(Global Information)
[CLS] Token Embedding
(Local Information)
Word Embeddings of the last four layers
MLP
MLP
Global Gated Attention
Feature Alignment Transformations1
Feature Alignment Transformations2
Weight1
Weight2
Cross Residual Transformer1
Cross Residual Transformer1
Cross Residual Transformer2
Cross Residual Transformer2
ReLU
GAP
O1
O2
O3
O4
Classifier
Global Image Embedding
Global Text Embedding
Global Gated Attention
MLP1
MLP2
S
Weight1
Weight2
Cross Residual Transformer
Add & Norm
Feed Forward
ConCat
Add & Norm
Add & Norm
Multi-Head Cross Residual Self-Attention
Input1
Input2
Feature Alignment Transformations1
1*1 ConV
BN
Flatten Dense
ReLU
LN
Feature Alignment Transformations2
Dense
ReLU
LN
Shared Weights
Element-wise Multiply
Concatenation Operation
Softmax Function
Scale
Dot Product
Sigmoid Function
Input1 Text / Image
Input2 Image / Text
MLP1 Fully-connected Layer
MLP2 Fully-connected Layer
GAP Global Average Pooling
BN Batch Normalization
LN Layer Normalization
Dense Fully-connected Layer

Fig. 2. The overview of our proposed HGLNet.

Image Encoder. Given a visual content V, we apply the ResNet50 [19] pre-trained on ImageNet dataset [20] to extract the region feature maps and global feature vector, which come from the "layer4" layer and the output layer of ResNet50, respectively. The processes can be represented as:

$$\mathbf{V_l} = \text{ResNet50}(V) = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}, \tag{2}$$

$$\mathbf{v_g} = \text{ResNet50}(V), \tag{3}$$

where $\mathbf{v}_n \in \mathbb{R}^{d_l}$ denotes the local region feature map, and $\mathbf{v}_g \in \mathbb{R}^{d_o}$ represents the global feature vector, respectively.

3.2. Multi-modal Feature Fusion

To effectively promote the fusion of textual and visual contents, we design a **G**lobal **G**ated **A**ttention (GGA) module and a **C**ross **R**esidual **T**ransformer (CRT) module to fuse the global and local features of each modality and obtain a discriminative fusion feature for classification.

Global gated attention module. Since the contribution of each modality to the final prediction is inconsistent, we consider this idea from a global perspective and design a GGA module for learning two weights to represent the contributions. The bottom left of Figure 2 presents the details of the module. Specifically, given two global feature vectors $\mathbf{v}_g$ and $\mathbf{e}_{\text{cls}}$, the vectors are first fed into two different fully-connected layers for dimensionality reduction, and two new vectors $\mathbf{v}_g \in \mathbb{R}^{d_g}$ and $\mathbf{e}_{\text{cls}} \in \mathbb{R}^{d_g}$ are obtained. Next, GGA concatenates them to get a new vector $\mathbf{f}_{\text{concat}}$ and then feeds it into two different fully-connected layers to output two vectors. Then, GGA applies the sigmoid functions to the original vectors and the obtained vectors, respectively. Finally, the dot products, scales, and softmax function are used to produce the weight_1 and weight_2. We summarize the operation of the GGA module in the following equations:

$$\mathbf{e}_{\text{cls}_1} = \sigma(\mathbf{f}_{\text{concat}}\mathbf{W}_1 + b_1), \mathbf{v}_{g_1} = \sigma(\mathbf{f}_{\text{concat}}\mathbf{W}_2 + b_2), \tag{4}$$

$$\text{weight}_1^{'} = \frac{\mathbf{e}_{\text{cls}} * \sigma(\mathbf{e}_{\text{cls}_1}^T)}{\sqrt{d_g}}, \text{weight}_2^{'} = \frac{\mathbf{v}_{g_1} * \sigma(\mathbf{v}_g^T)}{\sqrt{d_g}}, \tag{5}$$

$$\text{weight}_1 = \frac{exp^{\text{weight}_1^{'}}}{exp^{\text{weight}_1^{'}} + exp^{\text{weight}_2^{'}}}, \tag{6}$$

$$\text{weight}_2 = \frac{exp^{\text{weight}_2^{'}}}{exp^{\text{weight}_1^{'}} + exp^{\text{weight}_2^{'}}}, \tag{7}$$

where $\sigma(.)$ is the sigmoid function, and $\mathbf{W}_1 \in \mathbb{R}^{(d_g+d_g)\times d_g}$ and $\mathbf{W}_2 \in \mathbb{R}^{(d_g+d_g)\times d_g}$ the trainable parameters of the two fully-connected layers.

Feature alignment transformations. Since the dimensions of the obtained feature maps and the word embeddings are not identical, feature alignments should be performed first, transforming them to feature vectors of the same dimension to promote the next operations. The bottom right of Figure 2 presents the details of the transformations. Let $\mathbf{E}_l = \{\mathbf{e}_1, \dots, \mathbf{e}_m\}$ and $\mathbf{V}_l = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ to represent the word embeddings and image region embeddings. Specifically, we first employ a series of non-linear transformation operations to transform them. The details can be formulated as follows:

$$\mathbf{E}_l^{'} = \text{LN}(\text{ReLU}(\mathbf{E}_l\mathbf{W}_e)) \otimes \text{weight}_1, \tag{8}$$

$$\mathbf{V}_l^{'} = \text{LN}(\text{ReLU}(\text{Flatten}(\text{BN}((\text{Conv}_{1*1}(\mathbf{V}_l))\mathbf{W}_v)) \otimes \text{weight}_2, \tag{9}$$

where $\mathbf{E}_l^{'} \in \mathbb{R}^{d_n \times d_f}$ and $\mathbf{V}_l^{'} \in \mathbb{R}^{d_n \times d_f}$. In Equations 8 and 9, $\text{LN}(\cdot)$ and $\text{BN}(\cdot)$ denote the layer normalization and batch normalization, respectively. $\text{ReLU}(\cdot)$ represents the relu function, Flatten$(\cdot)$ denotes the flatten operation, and $\mathbf{W}_e \in \mathbb{R}^{d_t \times d_f}$ and $\mathbf{W}_v \in \mathbb{R}^{d_l \times d_f}$. Finally, we multiply the weight$_1$ and weight$_2$ with $\mathbf{E}_l'$ and $\mathbf{V}_l'$ element by element to obtain the matrices $\mathbf{E}_l' \in \mathbb{R}^{d_n \times d_f}$ and $\mathbf{V}_l' \in \mathbb{R}^{d_n \times d_f}$, respectively, and $\otimes$ denote the element-wise multiply.

Cross residual transformer. To better build the multi-modal semantic interaction information, we design a novel CRT module to fusion the local features of textual and visual modalities. The bottom left of Figure 2 presents the details of the module. CRT can capture the fine-grained semantic interactions and output a fusion feature for classification. Specifically, we propose a novel **D**ual-stream **C**ross **R**esidual **S**elf-**A**ttention (DCRSA) to replace the traditional **S**elf-**A**ttention (SA). As shown in Figure 3, our DCRSA module has two branches (I $\rightarrow$ T CRSA and T $\rightarrow$ I CRSA), and each branch focuses on capturing different semantic interactions.

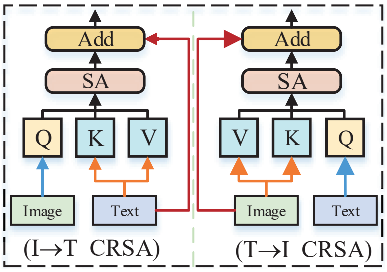

Fig. 3. The overview of our proposed CRSA.

Like SA, our DCRSA also receives a query $\mathbf{Q} \in \mathbb{R}^{d_n \times d_k}$, a key $\mathbf{K} \in \mathbb{R}^{d_n \times d_k}$, and a value $\mathbf{V} \in \mathbb{R}^{d_n \times d_v}$ as inputs and obtains a fusion feature. However, unlike SA, our **Q**, **K**, and **V** are from different modalities. Taking the T $\rightarrow$ I CRSA as an example, the inputs are word embeddings $\mathbf{E}_l'$ and image region embeddings $\mathbf{V}_l'$, respectively. First, we use $\mathbf{E}_l'$ to generate Q and $\mathbf{V}_l'$ to generate K, V, respectively. Then, T $\rightarrow$ I CRSA computes an inter-modality (Text to Image) matrix $\mathbf{F}_1$ as follows:

$$\mathbf{F}_1 = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)\mathbf{V} + \mathbf{V}_l', \tag{10}$$

where $\mathbf{F}_1 \in \mathbb{R}^{d_n \times d_v}$ represents the semantic interactions of an image to text, and softmax$(\cdot)$ denotes the softmax function. In Equation 10, we employ a residual connection to further obtain a better fusion feature. Similarly, I $\rightarrow$ T CRSA also computes an inter-modality (Image to Text) matrix $\mathbf{F}_2 \in \mathbb{R}^{d_n \times d_v}$. Note that $d_k = d_v = d_f$.

According to the description of the Transformer encoder, we can know that the Transformer encoder includes two sub-layers: the multi-head self-attention mechanism layer and the fully-connected feed-forward network. Meanwhile, residual connections and layer normalizations are applied. Therefore, as shown in Figure 2, we can define the **M**ulti-**H**ead **C**ross **R**esidual **S**elf-**A**ttention (MHCRSA) mechanism and the CRT:

$$M_1(Q,K,V) = \text{ConCat}(\mathbf{F}_{1,1},\ldots,\mathbf{F}_{1,h})\mathbf{W}_{O_1}, \tag{11}$$

$$M_2(Q,K,V) = \text{ConCat}(\mathbf{F}_{2,1},\ldots,\mathbf{F}_{2,h})\mathbf{W}_{O_2}, \tag{12}$$

$$\text{output}_1 = \text{LN}(\mathbf{E}_l' + M_1(Q,K,V)), \tag{13}$$

$$\text{output}_2 = \text{LN}(\mathbf{V}_l' + M_2(Q,K,V)), \tag{14}$$

$$\mathbf{O}_f = \text{ConCat}(\text{output}_1, \text{output}_2), \tag{15}$$

$$\mathbf{O}' = \text{LN}(\mathbf{O}_f + \text{FFN}(\mathbf{O}_f)), \tag{16}$$

where $M_1(Q,K,V)$ and $M_2(Q,K,V)$ denotes the MHCRSA functions and h represents the $h-th$ head. In Equations 11 and 12, the $\mathbf{W}_{O_1} \in \mathbb{R}^{(h \times d_v) \times d_f}$ and $\mathbf{W}_{O_2} \in \mathbb{R}^{(h \times d_v) \times d_f}$ are the parameter matrices of linear projections. In Equations 15 and 16, ConCat$(\cdot)$ denotes the concatenation operation, and FFN is a two-layer fully-connectd network that introduces non-linear transformation into the CRT. Moreover, $\mathbf{O}' \in \mathbb{R}^{d_n \times (d_f + d_f)}$ is the output of the CRT module.

3.3. Classification

$\mathbf{O}' \in \mathbb{R}^{d_n \times (d_f + d_f)}$ is the input of the classifier. Moreover, we know that BERT encodes the rich hierarchical semantic information of textual content [21]. Therefore, we apply the CRT on the outputs of the last four layers of BERT. However, this will bring a lot of parameters and increase the computational complexity of the HGLNet. To address the issue, we only apply 2 CRT modules to share their parameters for obtaining 4 different outputs. Furthermore, the outputs are fed into the relu function and global average pooling layer, which can denote $\mathbf{O}_1, \mathbf{O}_2, \mathbf{O}_3, \mathbf{O}_4$, respectively. Finally, we concatenate all outputs:

$$\mathbf{O} = \text{ConCat}(\mathbf{O}_1, \mathbf{O}_2, \mathbf{O}_3, \mathbf{O}_4), \tag{17}$$

where **O** is the final fusion feature representation of the multi-modal sample $S = \{T, V\}$ by the proposed HGLNet and GAP(.) denotes the global average pooling. Finally, **O** is fed into a fully-connected network with a softmax function to predict the label. Meanwhile, we apply cross-entropy as our loss function to train our network. The dropout is also employed during the training process to avoid model overfitting.

4. EXPERIMENTS

4.1. Settings

This section introduces the datasets, baselines, and implementation details.

Datasets. Three standards used benchmark datasets, i.e., WEIBO [8], Yelp [10], and CMCD-I [6] are used in multi-modal fake news detection, multi-modal sentiment analysis, and multi-modal tweet classification, respectively.

Baselines. For the multi-modal fake news detection task, we compare our network with four state-of-the-art models, MVAE [22], CARMN [11], HMCAN [3], and MCAN [12].

Table 1. Experiment results on three multi-modal tasks. (Unit:%)

Fake News Detection (Weibo Dataset)				Sentiment Analysis (Yelp Dataset)							Tweet Classification (CMCD-I Dataset)		
Methods	Accuracy	Fake	Real	Methods	BO	CH	LA	NY	SF	Avg.	Methods	Accuracy	F1
		F1	F1										
MVAE [22]	82.40	80.90	83.70	HAN-a [10]	55.18	54.88	53.11	52.96	51.98	53.16	ResNet	52.66	55.22
CARMN [11]	85.30	85.10	85.40	HAN-m [10]	56.77	57.02	55.06	54.66	53.69	55.01	BERT	65.28	65.85
HMCAN [3]	88.50	88.10	89.00	VistaNet [10]	63.81	65.74	**62.01**	61.08	60.14	61.88	ConCat	67.58	65.81
MCAN [12]	89.90	90.10	89.70	LD-MAN [23]	61.90	64.00	61.02	61.57	59.47	61.22	Fusion	76.14	75.82
HGLNet	**90.45**	**91.06**	**89.75**	HGLNet	**65.47**	**69.58**	60.78	**63.43**	**60.35**	**63.92**	HGLNet	**78.45**	**76.86**

We employ two baselines and VistaNet in [10], and LD-MAN in [23] to compare with our network for the multi-modal sentiment analysis task. Moreover, for the multi-modal tweet classification task, due to fewer baseline models, we compare our network with two single-modality models, ResNet [19] and BERT [18]; two fusion models, concatenation and fusion [6].

Implementation Details. Given the marked variations between the datasets, we adopt different hyperparameters and pre-process methods to train our network. We follow the previous work for the division of the datasets [3, 6, 10]. For the images in all datasets, we first resize images to $224 \times 224 \times 3$ and feed them to the pre-trained ResNet50 to obtain 1000-dimension feature vector (i.e., $d_o = 1000$ and feature maps $d_l = 2048 \times 7 \times 7$). For the text in all datasets, we first pre-process all the redundancy and noise (e.g., URLs, "@" symbols), and then we employ the pre-trained BERT to extract the 768-dimension token embedding of [CLS] and word embeddings (i.e., $d_t = 768$). We use AdamW optimizer [24] to optimize the loss function. We choose Accuracy and F1 as our evaluation metrics. In this paper, $d_g = 10$ and Dropout = 0.5. Moreover, the rest hyperparameters are shown in Table 2.

Table 2. Hyperparameters in the experiments.

Hyperparameter	Datasets		
	Weibo	Yelp	CMCD-I
BERT version	base_chinese	base_uncased	base_chinese
Text max_length (d_n)	140	300	300
$d_f = d_k = d_v$	32	64	64
Heads: h	4	8	4
Initial learning rate	0.0001	0.0002	0.0001
Weight decay	0.01	0.1	0.1
Minibatch size	128	160	180
Epochs	30	40	40

4.2. Performance Evaluation and Results Analysis

Table 1 reports the experimental results. HGLNet achieves the best performance for multi-modal fake news detection. It outperforms the existing state-of-the-art methods, which demonstrates that HGLNet can fuse multi-modal features better and jointly model multi-modal hierarchical information for fake news detection. In addition, CARMN, HMCAN, and MCAN all apply the transformer architecture, which proves the effectiveness of SA in multi-modal fusion. For the multi-modal sentiment analysis, we evaluate the proposed HGLNet on the Yelp dataset used in [10,23]. Since the dataset includes a rating scale of 1 to 5 as sentiment labels, we treat each rating as a sentiment class. The test datasets covered five cities, including Boston (BO), Chicago (CH), LosAngeles (LA), New York (NK), and San Francisco (SF). We observe that HGLNet achieves comparable performance for multi-modal sentiment analysis. For multi-modal tweet classification, since CMCD-I collects tweets with text and images from 18 general categories (e.g., game, finance), we perform single-modal and multi-modal experiments. We observe that textual or visual modality performance is worse than multi-modality, which confirms that multi-modal fusion can enhance classification performance. Moreover, HGLNet outperforms the concatenation fusion, which shows that a simple concatenation cannot model the multi-modal fusion better.

4.3. Ablation Study

To evaluate the effectiveness of the proposed GGA and CRT modules, we conduct ablation experiments on the Weibo dataset. We remove textual modality, visual modality, hierarchical semantics, and each module from the entire model for comparison, respectively. Table 3 displays the experimental results. In Tabel 3, "HGLNet" denotes the whole model with all modules, including textual semantics (T), visual semantics (V), hierarchical semantics (H), GGA module (G), and CRT module (C). After removing each component of HGLNet, we obtain the sub-models "-T", "-V", "-H", "-G", and "-C", respectively. We can observe that each module plays an efficient role in improving the performance of HGLNet. Specifically, the result of HGLNet-V is different from HGLNet-T, and HGLNet beats HGLNet-G, which reveals the contribution of textual and visual modality for the final classification is inconsistent. Moreover, the HGLNet-H and HGLNet-C are worse than HGLNet, which shows that our CRT can capture the multi-modal semantic interactions, and fusing hierarchical semantics can improve the performance of HGLNet.

Table 3. Experiment results of HGLNet ablation analysis.

HGLNet ablation analysis in Accuracy and F1 (Unit:%)			
Methods	Accuracy	Fake News	Real news
		F1	F1
HGLNet-T	80.90	80.20	81.50
HGLNet-V	74.00	74.40	73.50
HGLNet-H	88.54	88.75	88.32
HGLNet-C	89.21	90.45	87.67
HGLNet-G	89.93	90.56	89.38
HGLNet	**90.45**	**91.06**	**89.75**

5. CONCLUSION

This paper proposes a generic network suitable for multi-modal classification tasks called **H**ierarchical **G**lobal-**L**ocal feature fusion **N**etwork (HGLNet). To represent the contribution of each modality, a GGA module is proposed to adaptively generate weights by fusing global features of textual and visual modalities, which can represent the contributions. Besides, the CRT module is used to capture the fine-grained local semantic interactions between two modalities. Furthermore, HGLNet utilizes hierarchical semantic information to generate fusion features. Experimental results show that compared with the state-of-the-art, our HGLNet has better performance in multi-modal fake news detection, multi-modal sentiment analysis, and multi-modal tweet classification.

6. REFERENCES

[1] Douwe Kiela, Edouard Grave, Armand Joulin, and Tomas Mikolov, "Efficient large-scale multi-modal classification," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[2] Weiyao Wang, Du Tran, and Matt Feiszli, "What makes training multi-modal classification networks hard?," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12695–12705.

[3] Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu, "Hierarchical multi-modal contextual attention network for fake news detection," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 153–162.

[4] Xiaocui Yang, Shi Feng, Daling Wang, and Yifei Zhang, "Image-text multimodal emotion classification via multi-view attentional network," *IEEE Transactions on Multimedia*, 2020.

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

[6] Yong Hu, Heyan Huang, Anfan Chen, and Xian-Ling Mao, "A cross-modal classification dataset on social network," in *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer, 2020, pp. 697–709.

[7] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[8] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 795–816.

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

[10] Quoc-Tuan Truong and Hady W Lauw, "Vistanet: Visual aspect attention network for multimodal sentiment analysis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 305–312.

[11] Chenguang Song, Nianwen Ning, Yunlei Zhang, and Bin Wu, "A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks," *Information Processing & Management*, vol. 58, no. 1, pp. 102437, 2021.

[12] Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu, "Multimodal fusion with co-attention networks for fake news detection," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 2560–2569.

[13] Ziqi Yuan, Wei Li, Hua Xu, and Wenmeng Yu, "Transformer-based feature reconstruction network for robust multimodal sentiment analysis," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4400–4407.

[14] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD explorations newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[15] Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian, "Novel visual and statistical image features for microblogs news verification," *IEEE transactions on multimedia*, vol. 19, no. 3, pp. 598–608, 2016.

[16] Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang, "Rumor detection on social media with bi-directional graph convolutional networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 549–556.

[17] Nan Xu, Wenji Mao, and Guandan Chen, "A co-memory network for multimodal sentiment analysis," in *The 41st international ACM SIGIR conference on research & development in information retrieval*, 2018, pp. 929–932.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

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

[20] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[21] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah, "What does bert learn about the structure of language?," in *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[22] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma, "Mvae: Multimodal variational autoencoder for fake news detection," in *The world wide web conference*, 2019, pp. 2915–2921.

[23] Wenya Guo, Ying Zhang, Xiangrui Cai, Lei Meng, Jufeng Yang, and Xiaojie Yuan, "Ld-man: Layout-driven multimodal attention network for online news sentiment recognition," *IEEE Transactions on Multimedia*, vol. 23, pp. 1785–1798, 2020.

[24] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2018.