D^2R: Dual-Branch Dynamic Routing Network for Multimodal Sentiment Detection

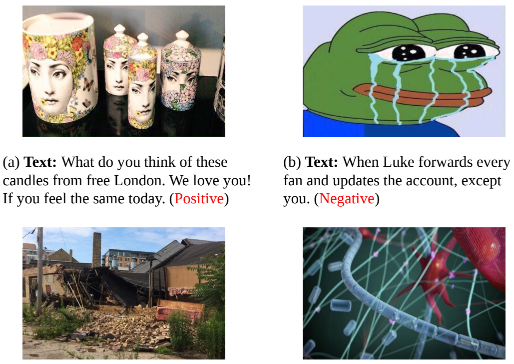
Yifan Chen[1*], Kuntao Li[1*], Weixing Mai[1], Qiaofeng Wu[1]
Yun Xue[1†], Fenghuan Li[2†]

[1]Guangdong Provincial Key Laboratory of Quantum Engineering and Quantum Materials,
School of Electronics and Information Engineering, South China Normal University
[2]School of Computer Science and Technology, Guangdong University of Technology

{chenyifan,likuntao,maiwx,scnu_wqf,xueyun}@m.scnu.edu.cn   fhli20180910@gdut.edu.cn

Abstract

Multimodal sentiment detection aims to classify the sentiment polarity of a given image-text pair. Existing approaches apply the same fixed framework to all input samples, lacking the flexibility to adapt to different image-text pairs. Furthermore, the interaction patterns of these methods are overly homogenized, limiting the model's capacity to extract multimodal sentiment information effectively. In this paper, we develop a **D**ual-Branch **D**ynamic **R**outing Network (**D^2R**), which is the first multimodal dynamic interaction model towards multimodal sentiment detection. Specifically, we design six independent units to simulate inter- and intra-modal information interactions without depending on any existing fixed frameworks. Additionally, we configure a soft router in each unit to guide path generation and introduce the path regularization term to optimize these inference paths. Comprehensive experiments on three publicly available datasets demonstrate the superiority of our proposed model over state-of-the-art methods.



(a) **Text:** What do you think of these candles from free London. We love you! If you feel the same today. (Positive)

(b) **Text:** When Luke forwards every fan and updates the account, except you. (Negative)

(c) **Text:** Ridge Avenue is closed after a partial building collapse and electrical fire Saturday night. (Negative)

(d) **Text:** This gel allows you to stop bleeding immediately: hemostatic drugs first aid. (Neutral)

Figure 1: Examples of multimodal sentiment tweets.

1 Introduction

With the growth of the Internet, people increasingly post multimodal messages on social media platforms to share opinions and express emotions (Yue et al., 2019; Mai et al., 2024; Chen et al., 2024). Consequently, multimodal sentiment detection has attracted significant attention in recent academic and industrial research, proving beneficial for tasks such as product review analysis and political opinion mining (Liang et al., 2021; Zhang et al., 2023). Unlike unimodal data, multimodal data provides richer information to reveal a person's true emotions (Zhang et al., 2018; Liang et al., 2021; Zhong et al., 2024).

In this work, we focus on multimodal sentiment detection for image-text pairs in social media posts. Previous works employed various fusion strategies to integrate features from different modalities (Xu, 2017; Xu and Mao, 2017). Other approaches introduced memory networks for inter-modal interactions (Xu et al., 2018; Yang et al., 2020). Yang et al. (2021) constructed a multi-channel graph neural network model for multimodal sentiment detection, Wei et al. (2023) recently utilized a sparse attention mechanism to enhance fusion by addressing modal heterogeneity. Although these methods have shown promising results, they predominantly applied existing static networks to handle all samples, with the fixed structure resulting in a lack of flexibility to adapt to different multimodal inputs, such as attention-based (Yang et al., 2020) and graph-based approaches (Yang et al., 2021). Moreover, existing related works are relatively homogenised in terms of interaction, focusing on capturing sentiment information through cross-modal alignment. In some cases, complex interaction patterns are unnecessary for simple image-text pairs and may introduce noise into the model (Qu et al., 2021). As shown in Fig. 1 (a), the word "love" in the text clearly indicates a "Positive" sentiment, making extensive interaction with the image redundant. Similarly, in

[*]These authors contributed equally to this work.
[†]Corresponding author.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 3536–3547
November 12-16, 2024 ©2024 Association for Computational Linguistics

Fig. 1 (b), the image depicts a crying frog, which directly conveys a "Negative" sentiment, reducing the need for textual interaction. Conversely, in other instances, varying levels of interaction between modalities are essential. In Fig. 1 (c), a local interaction between the" broken house" in the image and "building collapse" in the text helps identify it as a "Negative" example. However, in Fig. 1 (d), no affective cues are captured when interacting the cross-modal global and local information, so the model classifies it as a neutral example.

To tackle the above problems, we propose a novel **D**ual-Branch **D**ynamic **R**outing Network (**D²R**) for multimodal sentiment detection, which is a fully dynamic neural network. Specifically, to effectively address complex multimodal sentiment posts, we design six distinct units to implement interaction operations under various scenarios for both text and image without depending on any existing fixed frameworks. Each unit is configured with a soft router to generate inference paths. Subsequently, we stack these six units in width and depth to construct a complete routing space, enabling the exploration of more complex interaction patterns to flexibly adapt to diverse multimodal inputs. We also design a path regularization term to measure the sentiment-path similarity among samples, aiming to optimize these inference paths. Finally, we perform block fusion (Ben-Younes et al., 2019) on the multimodal inference information from different branches and feed it into the classifier for sentiment classification.

The main contributions of this paper are summarized as follows:

- We propose a novel dual-branch dynamic routing network that dynamically selects routing paths for diverse image-text pairs. To the best of our knowledge, we are the first to utilize a dynamic routing network to capture affective cues from different modalities for multimodal sentiment detection.

- We design six independent units to simulate inter- and intra-modal information interactions, with each unit integrating a soft router for routing learning and inference path optimization through path regularization.

- We conduct extensive experiments on three public datasets, and the results demonstrate the effectiveness of our model.

2 Related Work

2.1 Multimodal Sentiment Detection

Early works on multimodal sentiment detection utilized CNN and LSTM to extract and fuse feature representations from different modalities (Xu and Mao, 2017; Xu, 2017). Xu et al. (2018) introduced CoMN, a co-memory network iteratively modeled cross-modal interactions. Yang et al. (2020) proposed MVAN, which stacked pool module-tuned memory network to fuse multimodal features. Additionally, Yang et al. (2021) developed MGNNS, a multichannel graph neural network to capture emotions from entire dataset. Recently, Li et al. (2022) proposed CLMLF, a model combined contrastive learning and multilayer fusion. Another work (Wei et al., 2023) introduced the modal heterogeneity and proposed a multiview calibration network to resolve inherent differences in modalities. Despite promising results, these networks relied on fixed frameworks with static mechanisms to capture affective cues, limiting their capacity to dynamically handle diverse multimodal sentiment posts. In contrast, we aim to develop a dynamic neural network (Han et al., 2023; Li et al., 2024) to process complex image-text pairs adaptively, thereby enhancing the performance of sentiment classification.

2.2 Dynamic Neural Networks

Unlike common static neural networks, the inference process of dynamic neural networks adjusts dynamically based on different samples (Qu et al., 2021). Early works on dynamic networks focused on updating model parameters dynamically (Perez et al., 2018; Veit and Belongie, 2018). Subsequent research aimed to design dynamic models that enable automatic tuning of network depth or width (Liu et al., 2017). Dynamic neural networks have also shown excellent performance in recent multimodal tasks (Han et al., 2023; Zhu et al., 2023; Qu et al., 2021). Qu et al. (2021) first applied routing mechanisms to the domain of image-text retrieval, Zhou et al. (2021) introduced TRAR, a Transformer-based model to dynamically schedule global and local dependencies for VQA. However, TRAR only performed routing on unimodal data. Tian et al. (2023) proposed DynRT-Net, a dynamic routing converter network for multimodal sarcasm detection, which activated different modules through hierarchical collaboration. The work only achieved local dynamics by modifying sectional frames within the Transformer. To the best of our knowledge, the application of dynamic mechanisms in multimodal sentiment detection has never been explored. Unlike previous related works, we design six interaction units without relying on any existing fixed frameworks to model various interaction scenarios and employ dynamic routing mechanisms to explore novel interaction patterns, achieving truly global dynamics.

3 Methodology

3.1 Modal-specific Encoder

Given the input $x = (x^t, x^v)$, where x^t and x^v denote the text and image. In this work, $x^t = \{s_i\}_{i=1}^{P}$, P is the length of the text, we use the pre-trained BERT model to generate the final word embedding $e_i^t \in \mathbb{R}^D$, i refers to the i-th word. D is the hidden dimension. The local text feature denotes as $e^t \in \mathbb{R}^{P\times D}$. For each image x^v, we first divide each image into K patches and then use the pre-trained ViT model to generate the final region embedding $e_j^v \in \mathbb{R}^D$, j refers to the j-th region. The local image feature denotes as $e^v \in \mathbb{R}^{K\times D}$. We also use $[CLS]$ token representation to get the global feature $\bar{e_t}$ and $\bar{e_v}$ for text and image.

3.2 Dual-Branch Dynamic Sentiment Interaction Module

To capture complex and diverse sentiment information in multimodal posts, we design six independent units to realize inter- and intra-modal sentiment information interaction. These units incorporate existing interaction patterns and can explore more unexcavated ones based on routing strategies, endowing our model excellent ability to understand emotions and reason sentiment. Formally, the six units can be summarized as follows:

$$S_m^{(n)} = \{ \begin{array}{c} H_m^{(n)}(X_m^{(n)}), m = 1 \text{ or } 2 \\ H_m^{(n)}(X_m^{(n)}, Y), m = 3,4,5,6 \end{array} \tag{1}$$

where $S_m^{(n)} \in \mathbb{R}^{M\times D}$ denotes the output of the m-th unit in the n-th layer. $H_m^{(n)}$ represents the interaction function of the m-th unit in the n-th layer. $X_m^{(n)} \in \mathbb{R}^{M\times D}$ is the local input feature of the m-th unit in the n-th layer, $Y \in \mathbb{R}^{N\times D}$ denotes the local input feature from other modality of the m-th unit in the n-th layer.

In this work, we implement two single symmetrical interactive branches. Specifically, in text-image (T2V) branch, we set $X = e^t$ (M=P) and $Y = e^v$ (N=K), as for image-text (V2T) branch, $X = e^v$ (M=K) and $Y = e^t$ (N=P).

In this section, we take the T2V branch as an example to detail these six independent units.

Simplified Sentiment-Semantic Rectifying Unit. For a simple image or a short sentence, human can judge its sentiment polarity at a glance, and complex interactions are unnecessary. Therefore, we design a rectifiable unit to simplify the original sentiment information. It can be formulated as: $H_1^{(n)}(e^t) = \mathrm{ReLU}(e^t)$.

Unimodal Sentiment-Semantic Reasoning Unit. There may exist sentiment and semantic similarities between the local fragments (different words or visual regions), so we design a USSR unit to capture these semantic dependencies. Specifically, we employ a multi-head self-attention mechanism to capture intra-modal fine-grained sentimental associations in different subspaces, as follows:

$$h_i = Attention(Q_i, K_i, V_i) \tag{2}$$

where h_i denotes the output of the i-th head, $Q_i, K_i \in \mathbb{R}^{n\times \frac{D}{h}}$, $V_i \in \mathbb{R}^{n\times \frac{D}{h}}$ denote the *query*, *key* and *value* of i-th head, respectively.

Then, we concatenate all the R heads:

$$O = Multihead(e^t) = Concat(h_1, \cdots, h_R) + e^t \tag{3}$$

Based on the above processes, our USSR unit can be summarized as:

$$H_2^{(n)}(e^t) = FFN(O) + O \tag{4}$$

where FFN represents a feed-forward layer with $ReLU$ activation function.

Cross-modal Local Sentiment-Semantic Matching Unit. To explore the correlations between inter-modal local segments to enrich the fine-grained affective information representations, a CLSSM unit is designed. Specifically, we first calculate the attention weight between fragments of divergent modalities as follows:

$$\omega_{i,j} = \frac{\exp(\lambda a_{ij})}{\sum_{j=1}^{K}\exp(\lambda a_{ij})}, w_i^v = \sum_{j=1}^{K} w_{i,j} e_j^v \tag{5}$$

where λ denotes the inversed temperature factor and a_{ij} denotes the cosine similarity between e_i^t and e_j^v. $\omega_{i,j}$ is the attention weight matrix. w_i^v refers to the attended visual context vector with respect to i-th word. The complete local attended visual context vector with respect to all words can be denoted as w^v.

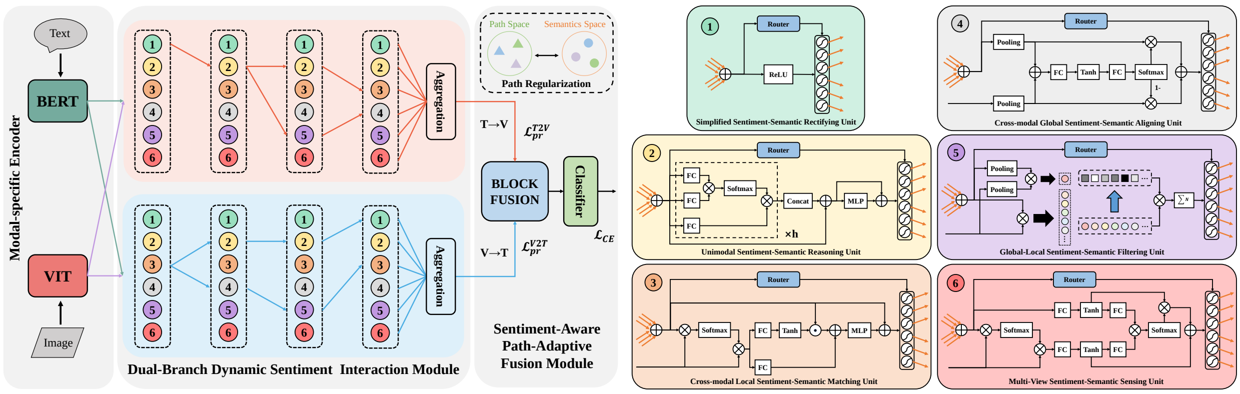


Figure 2: The overall architecture of the proposed D^{2}R model.

Then, we map w_i^v to generate the scaling vector α_i^v and the shifting vector β_i^v as follows:

$$\alpha_i^v = Tanh(FC_\alpha(w_i^v)) \tag{6}$$

$$\beta_i^v = FC_\beta(w_i^v) \tag{7}$$

Next, we take affine transformation operation followed by a MLP and residual connection to get the refined local fragment representation $\tilde{e_i^t}$ as:

$$\tilde{e_i^t} = MLP(e_i^t \odot \alpha_i^v + \beta_i^v) + e_i^t \tag{8}$$

where $\odot$ denotes element-wise multiplication.

Combining the above steps, our CLSSM unit can be summarized as: $H_3^{(n)}(e_i^t, e_j^v) = [\tilde{e_1^t}; \cdots; \tilde{e_P^t}]$.

Cross-modal Global Sentiment-Semantic Aligning Unit. Compared with the local fragments information, the global holistic information can reflect the overall sentiment of one text-image pair on a broader level, so we design a CGSSA unit to integrate the global sentiment information of different modalities to learn valuable cross-modal coarse-grained representations. Specifically, we introduce a special gated fusion mechanism to adaptively combine the global text representation $\bar{e}_t$ and visual representation $\bar{e}_v$, which is formulated as:

$$\bar{e} = z\bar{e}_t + (1-z)\bar{e}_v \tag{9}$$

$$z = \frac{\exp(W_1\delta(W_2\bar{e}_t))}{\exp(W_1\delta(W_2\bar{e}_t)) + \exp(W_1\delta(W_2\bar{e}_v))} \tag{10}$$

where W_1 and $W_2 \in \mathbb{R}^{m\times d}$ are parameter matrices, δ denotes the Tanh function.

After the above processes, our CGSSA unit can be profiled as: $H_4^{(n)}(\bar{e}_t, \bar{e}_v) = \bar{e}$.

Global-Local Sentiment-Semantic Filtering Unit. For complex text-image pairs, relying only on cross-modal global or local information is still insufficient to classify sentiment. Therefore, we design a GLSSF unit to capture both the cross-modal fine-grained and coarse-grained emotional cues simultaneously, and compensate for affective differences. In addition, we notice that indiscriminately aggregate all possible local comparisons and global comparisons may cause less-meaningful comparisons (such as "a" and "the" correlation comparisons), which hinder the model's capacity to distinguish sentiment polarity. Therefore, we deliberately develop a strategy in GLSSF unit to effectively suppress invalid comparisons with low affective contributions. Specifically, we first compute cross-modal global and local sentimental similarity vector (Diao et al., 2021) as follows:

$$F(a, b; W_f) = \frac{W_f|a-b|^2}{\|W_f|a-b|^2\|_2} \tag{11}$$

where $a, b \in \mathbb{R}^D$ are two different vectors, $|\cdot|^2$ and $\|\cdot\|^2$ separately represent the element-wise square and l_2-norm. $W_f \in \mathbb{R}^{m\times d}$ is a parameter matrix.

Thereafter, we compute the cross-modal global or local sentiment similarity as:

$$F^{Global} = F(\bar{e}_t, \bar{e}_v; W_f^g) \tag{12}$$

$$F_i^{Local} = F(e_i^t, w_i^v; W_f^l) \tag{13}$$

where $W_f^g, W_f^l \in \mathbb{R}^{m\times d}$ are parameter matrices.

Next, we calculate the aggregation weight γ_u for the obtained cross-modal global and each local sentiment similarity vector representations $N = \{F_1^{Local}, \cdots, F_i^{Local}, F^{Global}\}$.

$$\gamma_u = \frac{\sigma(BN(W_\gamma F_u))}{\sum_{F_v\in N}\sigma(BN(W_\gamma F_v))} \tag{14}$$

where σ denotes the sigmoid function, BN indicates the batch normalization, and $W_\gamma \in \mathbb{R}^{m\times 1}$ is a linear transformation.

Finally, we converge all the sentiment similarity representations as follows:

$$\gamma_f = \sum_{F_u \in N} \gamma_u F_u \tag{15}$$

Combining the above processes, our GLSSF unit can be represented as: $H_5^{(n)}(e^t, e^v, \bar{e_t}, \bar{e_v}) = \gamma_f$.

Multi-View Sentiment-Semantic Sensing Unit. Both unimodal and cross-mdoal information are beneficial for the fianl sentiment classification. Therefore, we design a MVSSS unit that learns unimodal context-rich cross-modal sentiment features from two different views, aiming to facilitate interactive reasoning between unimodal and multimodal sentiment information. Specifically, considering the possible common features between unimodal and cross-modal information, we first project w^v and e^t into a common potential semantic space for sentiment-semantic matching, as follows:

$$C_w = Tanh(W_w w^v + b_w) \tag{16}$$

$$C_e = Tanh(W_e e^t + b_e) \tag{17}$$

where C_w and C_e denote the converted multi-view cross-modal and unimodal sentiment features in same space.

Next, we modify the gating mechanism to filter possible sentimental differences noise to integrate the common features of unimodal and cross-modal information. And then we learn the unimodal context-rich cross-modal sentiment features. Specifically, we align C_w based on C_e, and set C_w as $Q_m = W_Q C_w$ and the C_e as $K_u = W_K C_e$, where W_Q and W_K are trainable parameters. Thus, the $V_m = softmax(Q_m K_u^T)$, where V_m is query attended mask. Thereafter, unimodal context-rich cross-modal sentiment feature C_{we} can be formulates as:

$$C_{we} = C_w + V_m C_e \tag{18}$$

After the above processes, our MVSSS unit can be summarized as: $H_6^n(e^t, w^v) = C_{we}$.

Soft Router. To fully utilize the inimitable strengths of six units, we set up the layers in parallel and connect them between adjacent layers in a dense manner. This dense connectivity ensures a multiple and flexible routing space where many unexcavated interaction patterns can be explored. After constructing the routing space, the routing process is executed by soft router, the input of the m-th unit in the n-th layer can be obtained by the following operation:

$$H_m^{(n)} = \begin{cases} e^t, & n = 0 \\ \sum_{j=0}^{c-1} \rho_{j,m}^{(n-1)} S_j^{(n-1)}, & n > 0 \end{cases} \tag{19}$$

where $C = 6$ indicates the total number of units in each layer. $S_j^{(n-1)} \in \mathbb{R}^{P\times D}$ represents the output of j-th unit in the $(n-1)$-th layer. $\rho_{j,m}^{(n-1)} \in [0, 1]$ is the path probability from the j-th unit in the $(n-1)$-th layer to the m-th unit in the n-th layer. This can be calculated as follows:

$$\rho_m^{(n)} = \mathrm{ReLU}\{\mathrm{Tanh}[MLP(\frac{1}{P}\sum_{r=1}^{P} h_{m,r}^{(n)})]\} \tag{20}$$

where $\rho_m^{(n)} \in \mathbb{R}^C$ denotes the path probability vector of all units in n-th layer, $h_{m,r}^{(n)}$ is the r-th row vector of $H_m^{(n)}$.

Thereafter, the routing process is finished, we can obtain the final refined feature matrix $H_{1-6}^* = H_{1-6}^{(L)}$ through Equation (20) from the last layer L. Then, we take average-pooling operation for aggregating the six units output embeddings H_{1-6}^* to obtain the final aggregated single-branch feature representation $\bar{h}_{1-6}$.

3.3 Sentiment-Aware Path-Adaptive Fusion Module

Block Fusion. We implement the units and routing process on text and image-modality respectively, and obtain two branches of aggregated feature representation, namely, $\bar{h}_{1-6}^{T2V}$ and $\bar{h}_{1-6}^{V2T}$. Then, we adopt a block fusion strategy to fuse $\bar{h}_{1-6}^{T2V}$ and $\bar{h}_{1-6}^{V2T}$ and use the fusion feature to make the final prediction. Inspired by Ben-Younes et al. (2019), we project $\bar{h}_{1-6}^{T2V}$ and $\bar{h}_{1-6}^{V2T}$ into a new feature space through the association tensor T, as follows:

$$f = T \times_1 \bar{h}_{1-6}^{T2V} \times_2 \bar{h}_{1-6}^{V2T} \tag{21}$$

where $\times_1$ and $\times_2$ means tensor product along different dimensional space.

The final fusion tensor f is fed into an additional MLP followed by softmax function to predict the sentiment label:

$$\hat{y} = softmax(MLP(f)) \tag{22}$$

where $\hat{y}$ denotes the prediction label.

Finally, we apply the cross-entropy loss function:

$$L_{CE} = -(ylog(\hat{y}) + (1-y)log(1-\hat{y})) \tag{23}$$

where y denotes the ground-truth label.

Path Regularization. In fact, the sentiment information and semantics of multimodal posts are key factors affecting the interaction patterns. Samples with similar sentiment polarity should learn similar routing paths, while samples with different sentiment polarity should learn discrepant routing paths as much as possible. We hope that the routing path distribution can be consistent with the sentiment-semantic distribution. Therefore, we introduce the path regularization term to measure their correlations among samples. Particularly, we take average-pooling on $e^t \in \mathbb{R}^{P \times D}$ to get the sentiment-semantic representation $\bar{e^t} \in \mathbb{R}^D$ and compute the sentiment-semantic similarity as $S_t = \bar{e^t} \cdot (\bar{e^t})^\top$. Thereafter, we connect the output values of all routers to get the path vector $\varepsilon^t \in \mathbb{R}^{C^{2(L-1)+C}}$ and compute the path similarity as $S_p = \varepsilon^t \cdot (\varepsilon^t)^\top$.

To achieve sentiment-path consistency, we develop a path regularization loss function L_{pr} to calculate the distribution gap between the sentiment-semantic representation S_t and the path vector S_p, which is formulated as:

$$L_{pr}^{T2V} = JS(S_t || S_p) \tag{24}$$

where JS stands for JS divergence (Sutter et al., 2020). Likewise, we can obtain the V2T branch loss function L_{pr}^{V2T}.

3.4 Training objective

The overall loss function for $\mathrm{D^2R}$ is as follows:

$$L_{All} = L_{CE} + \lambda_1 L_{pr}^{T2V} + \lambda_2 L_{pr}^{V2T} \tag{25}$$

where λ_1 and λ_2 control the ratio of L_{pr}^{T2V} and L_{pr}^{V2T}, respectively.

4 Experiments

4.1 Experiment Settings

Dataset. We assess our model by conducting experiments on three publicly available benchmark datasets which are **MVSA-Single**, **MVSA-Multiple** and **HFM**. The statistics of the dataset are shown in Appendix A.1.

Implementation. The details of parameter implementations are listed in Appendix A.2.

Baselines. We compare our model with unimodal baseline models and multimodal models.

Table 1: Experimental results of different models on MVSA-Single, MVSA-Multiple and HFM datasets.

	MVSA-Single		MVSA-Multiple			HFM	
Model	**ACC**	**F1**	**ACC**	**F1**	**Model**	**ACC**	**F1**
			Text-Only				
CNN	0.6819	0.5590	0.6564	0.5766	CNN	0.8003	0.7532
BiLSTM	0.7012	0.6506	0.6790	0.6790	BiLSTM	0.8190	0.7753
BERT	0.7111	0.6970	0.6759	0.6624	BERT	0.8389	0.8326
TGNN	0.7034	0.6594	0.6967	0.6180			
			Image-Only				
ResNet	0.6467	0.6155	0.6188	0.6098	ResNet	0.7277	0.7138
ViT	0.6378	0.6226	0.6194	0.6119	ViT	0.7309	0.7152
OSDA	0.6675	0.6651	0.6662	0.6623			
			Multi-Modal				
MultiSentiNet	0.6984	0.6984	0.6886	0.6811	Concat(2)	0.8103	0.7799
HSAN	0.6988	0.6690	0.6796	0.6776	Concat(3)	0.8174	0.7874
Co-MN-Hop6	0.7051	0.7001	0.6892	0.6883	MMSD	0.8344	0.8018
MGNNS	0.7377	0.7270	**0.7249**	0.6934	D&R Net	0.8402	0.8060
CLMLF	0.7533	0.7346	0.7200	0.6983	CLMLF	0.8543	0.8487
MVCN	0.7606	0.7455	0.7207	0.7001	MVCN	0.8568	0.8523
$\mathrm{D^2R}$	**0.7667**	**0.7559**	0.7159	**0.7085**	**$\mathrm{D^2R}$**	**0.8672**	**0.8625**

Unimodal Baselines. For text-modality, we choose **CNN**, **BiLSTM**, **BERT** and **TGNN** as baselines. For image-modality, **ResNet** , **ODSA** and **ViT** are three popular models.

Multimodal Baselines. For MVSA-Single and MVSA-Multiple datasets, including: **MultiSentiNet**, **HSAN**, **Co-MN-Hop6**, **MGNNS**, **CLMLF** and **MVCN**. For HFM dataset, including: **Concat(2)** and **Concat(3)**, **MMSD**, **D&R Net**. More details on baselines are provided in Appendix A.3.

4.2 Experiments results

We evaluate the effectiveness of our proposed framework by comparing it with the baseline models as shown in Table 1 and derive the following observations. 1). It is evident that both the text and image play crucial roles in sentiment detection. Therefore, it is imperative to fully excavate the affective cues from different modalities, which validate our tuition of designing two single symmetric branches of the two-channel interaction. In addition, the multi-modal models consistently outperform the unimodal models on performance because of fusing more sentiment information. 2). Our $\mathrm{D^2R}$ achieves considerable improvement on Acc and F1 compared with the other strong baseline models on the three datasets, which suggests that dynamic routing network have advantages over regular static networks. 3). At last, we find that $\mathrm{D^2R}$ achieves better results on HFM dataset compared to the MVSA datasets. The reason may be that for classification tasks with fewer label categories, the interaction patterns contained in the six units capture more accurate sentiment information.

Table 2: Ablation experiment results of our model.

	MVSA-Single		MVSA-Multiple		HFM	
Model	ACC	F1	ACC	F1	ACC	F1
$\mathrm{D^2R}$	**0.7667**	**0.7559**	**0.7159**	**0.7085**	**0.8672**	**0.8625**
w/o 1	0.7244	0.7234	0.7041	0.6931	0.8501	0.8438
w/o 2	0.7467	0.7465	0.7088	0.6880	0.8588	0.8529
w/o 3	0.7289	0.7155	0.7005	0.6823	0.8559	0.8500
w/o 4	0.7267	0.7284	0.6952	0.6661	0.8580	0.8539
w/o 5	0.7067	0.7082	0.6652	0.6661	0.8592	0.8543
w/o 6	0.7333	0.7208	0.6764	0.6702	0.8630	0.8582
w/o BF	0.7333	0.7265	0.6800	0.6707	0.8584	0.8537
w/o PR	0.7467	0.7482	0.7047	0.6986	0.8617	0.8570

Table 3: Soft router ablation study experiment results.

	MVSA-Single		MVSA-Multiple		HFM	
Model	ACC	F1	ACC	F1	ACC	F1
$\mathrm{D^2R}$	**0.7667**	**0.7559**	**0.7159**	**0.7085**	**0.8672**	**0.8625**
w/o Soft Router	0.7244	0.7199	0.6911	0.6799	0.8526	0.8470
Random Router	0.7067	0.7106	0.6894	0.6886	0.8517	0.8473
Hard Router	0.7378	0.7317	0.7082	0.7000	0.8567	0.8523

Table 4: Cosine similarity calculations ablation study experiment results.

	MVSA-Single		MVSA-Multiple		HFM	
Model	ACC	F1	ACC	F1	ACC	F1
$\mathrm{D^2R}$	**0.7667**	**0.7559**	**0.7159**	**0.7085**	**0.8672**	**0.8625**
Manhattan distance (L1)	0.7244	0.7232	0.6888	0.6778	0.8430	0.8395
Euclidean distance (L2)	0.7533	0.7441	0.7076	0.6985	0.8542	0.8492
Mean Squared Displacement (MSD)	0.7333	0.7292	0.7076	0.6922	0.8559	0.8510

4.3 Ablation Study

To further investigate the effectiveness of each component in $\mathrm{D^2R}$, we conduct a series of ablation studies: **1) w/o 1:** we remove the SSSR unit; **2) w/o 2:** we remove the USSR unit; **3) w/o 3:** we remove the CLSSM unit; **4) w/o 4:** we remove the CGSSA unit; **5) w/o 5:** we remove the GLSSF unit; **6) w/o 6:** we remove the MVSSS unit; **7) w/o BF:** we remove the block fusion module; **8) w/o PR:** we remove the path regularization term.

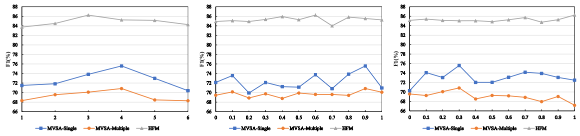
Table 2 shows the results of ablation study. It is evident that the performance after removing any of the components is worse than the original $\mathrm{D^2R}$, which demonstrates the effectiveness of each component. Specifically, for MVSA-Single and MVSA-Multiple datasets, w/o 5 degrades dramatically, it drops absolutely 0.0600 and 0.0507 on ACC, 0.0477 and 0.0424 on F1, respectively. This demonstrates that GLSSF unit can compensate for sentimental differences by capturing cross-modal global and local affective cues. At the same time, it verifies the rationality of calculating the sentiment similarity vector of cross-modal global and local information, suppressing irrelevant ones. For HFM dataset, w/o 1 has the most significant decline which indicates that the simplest SSSR unit plays an important role. We speculate that the reason may be that the HFM dataset has more simple text-image pairs. Moreover, w/o 6 achieve better results than others, only decrease 0.0042 on ACC and 0.0043 on F1 as HFM is a simple binary classification task dataset, the unit that capture sentiment information from complex multiple perspectives may play less important role. The performance of w/o BF declines distinctly. This suggests that block fusion benefits our dual-branch model by obtaining better sentiment representation. Particularly, for w/o PR, the ACC and F1 of the three datasets also show varying degrees of performance degradation. It proves the effectiveness of our proposed path regularization which consider the the consistency of routing path and sentiment semantics.

We also execute three additional ablation studies to verify the rationality of our proposed soft router, including: **1) w/o Soft Router:** we remove the soft router instead of selecting routing paths, **2) Random Router:** we replace the soft router with the random router, deriving the path probability of each unit from a uniform distribution, **3) Hard Router:** we replace the soft router with the hard router, introducing the gumbel-softmax trick to discretize path values. We report the experimental results in Table 3 and have the following observations: the metrics of $\mathrm{D^2R}$ significantly outperform other three methods on all datasets, demonstrating the effectiveness of our proposed soft router equivalent to other methods. Moreover, hard router got the lightest drop in performance compared with random router and w/o soft router, we hypothesize that the reason may be that random router may introduce extra path noise into the model, while not using router deprives the model's ability to dynamically adapt to different inputs.

To validate the reliability of the cosine similarity calculations using embeddings obtained from different pre-trained models in the cross-modal local sentiment-semantic matching unit, we design more three ablation studies to compared the effect of the similarity distance calculation formulas we used and others on the final model's performance. The experimental results are shown in Table 4. It is clear that using cosine values to compute the similarity of embeddings from different pre-trained models is more efficient than other methods and has the most significant improvement in model's performance. At the same time, some previous works have verified the advantages of using cosine similarity calculations (Diao et al., 2021; Chen et al., 2022; Zhang et al., 2022).

Table 5: Experimental results of model computational complexity.

Metric	MVSA-Single	MVSA-Multiple	HFM
Batch size	64	64	64
Model parameters	459.87M	459.87M	407.83M
FLOPs	18,221.84M	23,282.92M	20,855.04M
Inference time	192.65ms	230.35ms	164.60ms
Max Memory Reserved	19.14GB	21.14GB	19.03GB
Max Memory Allocated	7.12GB	7.05GB	6.32GB
GPU usage	NVIDIA 3090 GPU	NVIDIA 3090 GPU	NVIDIA 3090 GPU



(a) dynamic routing layers L (b) path regularization loss weight factor λ_1 (c) path regularization loss weight factor λ_2

Figure 3: The influence of hyper-parameters.

4.4 Hyperparameter Analysis

To analyze the impact of the number of dynamic routing layers L in our model, we conduct experiments on varying the layer of dynamic routing from 1 to 6. The results are shown in Figure 3. For MVSA-Single and MVSA-Multiple datasets, we can see that the performance metric F1 improves with the increase of dynamic routing layers in the range 1 to 4, and then drops slightly while the routing layers exceed 4. For HFM dataset, F1 improves with the increase of dynamic routing layers in the first 3 layers, and then decreases in the layers 4 to 6. The results show that increasing the number of routing layers in an appropriate range can improve the performance as more layers offer broader path space, thus increasing the ability of exploring more superior interaction patterns. However, when layers exceed 3 or 4, overfitting limits model optimization and hinders the path learning ability.

In addition, we also carry out several experiments on λ_1 and λ_2 to research the influence of the path regularization parameters L_{pr}^{T2V} and L_{pr}^{V2T} on the final prediction. The results are shown in Figure 3. For MVSA datasets, the performance first comes and goes before the saturation points $(\lambda_1 = 0.9, \lambda_2 = 0.3)$, and then begins to decline when λ_1 exceed 0.9 and λ_2 exceed 0.3. We can infer that the saturation points $(\lambda_1 = 0.9, \lambda_2 = 0.3)$ can maximize the similarity of dynamic paths for examples with the same sentiment polarity. For HFM, the best λ_1 is 0.6 and λ_2 is 1.0. Then, a slight drop in performance occurs on other values. Apparently, excessively large λ_1 and λ_2 affect the performance on three datasets, the reason may be that overly exploring the path diversity leads to a terrible over-fitting phenomenon.

4.5 Computational Complexity Analysis

With the introduction of units and more paths, we perform complementary experiments to analyze the model's computational complexity, inference latency and parameter count. We report the model parameters, FLOPs, inference time, max memory reserved and max memory allocated for the model on three different datasets in Table 5. FLOPs refer to the floating point operations, which are used to evaluate the model's computational complexity; inference time reflects the model's inference latency; model parameters reflect the model's total parameter count. On both MVSA datasets, the number of dynamic routing layers is set to 4, on HFM dataset, it is set to 3. Thus our model has more inference paths and parameter count on MVSA datasets than HFM dataset.

4.6 Visualization

To demonstrate the vital advantage of our dynamic reasoning sentiment methods. We thus show some images and visualize the path vectors learned in SAPAF module, which indicate that D^{2}R can adaptively choose the best paths for different examples. Specifically, we use the t-SNE (Van der Maaten and Hinton, 2008) algorithm to map the concatenation path vector into a 2-dimensional Euclid space. Afterwards, we clustered these 2-dimensional vectors into 6 groups in different 6 colors. As shown in Figure 4, we could observe that the images related to obvious positive emotions (the points marked in brown and yellow) and the ones related to obvious negative emotions (the points marked in blue and green) can be well distinguished. For instance, there exists a large margin between brown points (associated with happy crowds) and blue points (associated with bad weather), because not only is there a semantic gap between "crowds" and "weather", but there is also a sentimental conflict between "happy" and "bad". Although both brown and green points are related to people, the sentiment differences between them are still significant, as a result, they are still wide apart in a 2-dimensional space. Besides, neutral examples serve as a demarcation between positive and negative examples and are located between the two. Pictures with no emotional inclination (the points marked in red and purple) are also can be well distinguished. Because the advanced fine-grained semantics is much different between red points (re-

Figure 4: Visualization of the learned path vectors.

lated to logos) and purple points (related to foods). Our proposed soft router can make path choices based on these fine-grained semantic and sentiment information to the path selection. These results reveal that our D^2R is able to adaptively learn specific semantic-related and sentiment-aware paths for diverse inputs, thus the distribution of learning paths is to a certain extent consistent with that of sentiment-semantic.

4.7 Case study

To further verify the adaptability of D^2R, we qualitatively visualize the routing process for several typical examples. As shown in Figure 5, we have the following observations: 1) Simpler text-image pairs tend to activate less paths as their sentiment polarity is obvious. For example, the sentence in Fig. 5 (a) conveys a distinctly positive emotion (love), which may not require much image information; 2) Cross-modal global content analysis can activate more paths for some examples to obtain accurate affective cues. In Fig. 5 (b), We can't tell the sentiment polarity of the post from only text-modality (image-modality), but when we focus on the entire text-image pairs, we can see that it's a negative example (a busted pen soiled hands like doing actual manual labor and working on the car). 3) Additional attention to fine-grained cross-modal local information can activate more paths to capture nuanced sentiment information. Fig. 5 (c) shows two intact strawberries and one damaged strawberry. We first understand the meaning of the entire text, and then paying extra attention to the comparison between "destroyed" in text and "damaged strawberry" in image to accurately classify it as a negative example. 4) The elements in the first three examples are relatively single (book, hand or food), their interaction patterns and routing paths are less complex than the fourth, because there exist many elements in Fig. 5 (d). ("bule sky", "white clouds", "green trees" ,"high mountains" and "exquisite tables"). We need to focus on all the elements one by one against the text, and then analyze the entire text and image to determine the final sentiment polarity.

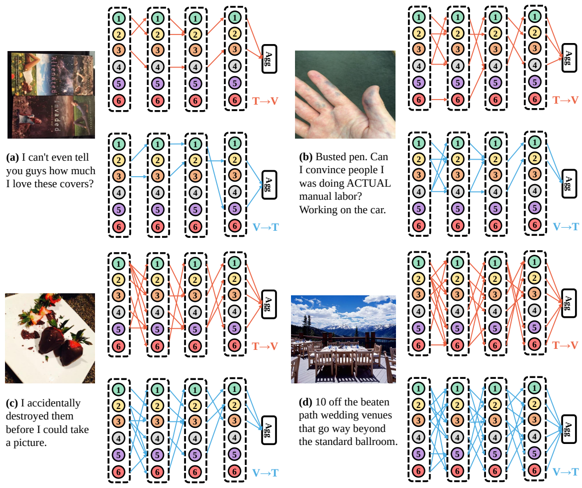


Figure 5: Path visualisation of interaction patterns for four typical cases.

5 Conclusion

This paper presents a novel dynamic neural network model for multimodal sentiment detection called D^2R, which is the first work on exploring diverse interaction patterns using dynamic routing mechanisms. Specifically, we apply six units to simulate various levels of inter- and intra-modal interaction patterns. A soft router is integrated to adapt flexibly to diverse image-text pairs through routing path learning. Additionally, we introduce a path regularization term to measure sentiment-path similarity between samples and optimize the inference path. Comprehensive experiments demonstrate that our model achieves state-of-the-art performance on three benchmark datasets.

Limitations

At this stage, we concentrate on two limitations of this work, aiming to inspire future potential research directions.

- For multimodal sentiment analysis of social media posts, incorporating more external knowledge to enrich sentiment semantic information could improve the model's predictive performance. Our model ignores the importance of external knowledge for this task.

- Multimodal dynamic routing networks can be extended to other multimodal tasks on social media, representing a primary focus for our future research.

Acknowledgement

This work was supported in part by Guangdong Basic and Applied Basic Research Foundation(2023A1515011370), National Natural Science Foundation of China (32371114), Characteristic Innovation Projects of Guangdong Colleges and Universities (2018KTSCX049), International Scientific and Technological Cooperation Project of Huangpu and Development Districts in Guangzhou (2022GH01, 2023GH05).

References

Hedi Ben-Younes, Remi Cadene, Nicolas Thome, and Matthieu Cord. 2019. Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8102–8109.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 2506–2515.

Jianan Chen, Lu Zhang, Qiong Wang, Cong Bai, and Kidiyo Kpalma. 2022. Intra-modal constraint loss for image-text retrieval. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 4023–4027. IEEE.

Yifan Chen, Haoliang Xiong, Kuntao Li, Weixing Mai, Yun Xue, Qianhua Cai, and Fenghuan Li. 2024. Relevance-aware visual entity filter network for multimodal aspect-based sentiment analysis. *International Journal of Machine Learning and Cybernetics*, pages 1–14.

Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. 2021. Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 1218–1226.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*.

Yudong Han, Jianhua Yin, Jianlong Wu, Yinwei Wei, and Liqiang Nie. 2023. Semantic-aware modular capsule routing for visual question answering. *IEEE Transactions on Image Processing*.

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

Lianzhe Huang, Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2019. Text level graph neural network for text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3444–3450, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the naacL-HLT*, volume 1, page 2.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Xingye Li, Jin Liu, Yurong Xie, Peizhu Gong, Xiliang Zhang, and Huihua He. 2024. Magdra: a multi-modal attention graph network with dynamic routing-by-agreement for multi-label emotion recognition. *Knowledge-Based Systems*, 283:111126.

Zhen Li, Bing Xu, Conghui Zhu, and Tiejun Zhao. 2022. CLMLF:a contrastive learning and multi-layer fusion method for multimodal sentiment detection. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2282–2294, Seattle, United States. Association for Computational Linguistics.

Bin Liang, Chenwei Lou, Xiang Li, Lin Gui, Min Yang, and Ruifeng Xu. 2021. Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs. In *Proceedings of the 29th ACM international conference on multimedia*, pages 4707–4715.

Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huang, Shoumeng Yan, and Changshui Zhang. 2017. Learning efficient convolutional networks through network slimming. In *Proceedings of the IEEE international conference on computer vision*, pages 2736–2744.

Weixing Mai, Zhengxuan Zhang, Yifan Chen, Kuntao Li, and Yun Xue. 2024. Geda: Improving training data with large language models for aspect sentiment triplet extraction. *Knowledge-Based Systems*, 301:112289.

Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El Saddik. 2016. Sentiment analysis on multi-view social data. In *MultiMedia Modeling: 22nd International Conference, MMM 2016, Miami, FL, USA, January 4-6, 2016, Proceedings, Part II 22*, pages 15–27. Springer.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Leigang Qu, Meng Liu, Jianlong Wu, Zan Gao, and Liqiang Nie. 2021. Dynamic modality interaction modeling for image-text retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1104–1113.

Rossano Schifanella, Paloma De Juan, Joel Tetreault, and Liangliang Cao. 2016. Detecting sarcasm in multimodal social platforms. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1136–1145.

Thomas Sutter, Imant Daunhawer, and Julia Vogt. 2020. Multimodal generative learning utilizing jensen-shannon-divergence. *Advances in neural information processing systems*, 33:6100–6110.

Yuan Tian, Nan Xu, Ruike Zhang, and Wenji Mao. 2023. Dynamic routing transformer network for multimodal sarcasm detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2468–2480, Toronto, Canada. Association for Computational Linguistics.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Andreas Veit and Serge Belongie. 2018. Convolutional networks with adaptive inference graphs. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–18.

Yiwei Wei, Shaozu Yuan, Ruosong Yang, Lei Shen, Zhangmeizhi Li, Longbiao Wang, and Meng Chen. 2023. Tackling modality heterogeneity with multi-view calibration network for multimodal sentiment detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5240–5252.

Nan Xu. 2017. Analyzing multimodal public sentiment based on hierarchical semantic attentional network. In *2017 IEEE international conference on intelligence and security informatics (ISI)*, pages 152–154. IEEE.

Nan Xu and Wenji Mao. 2017. Multisentinet: A deep semantic network for multimodal sentiment analysis. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2399–2402.

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

Nan Xu, Zhixiong Zeng, and Wenji Mao. 2020. Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3777–3786.

Xiaocui Yang, Shi Feng, Daling Wang, and Yifei Zhang. 2020. Image-text multimodal emotion classification via multi-view attentional network. *IEEE Transactions on Multimedia*, 23:4014–4026.

Xiaocui Yang, Shi Feng, Yifei Zhang, and Daling Wang. 2021. Multimodal sentiment detection based on multi-channel graph neural networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 328–339.

Lin Yue, Weitong Chen, Xue Li, Wanli Zuo, and Minghao Yin. 2019. A survey of sentiment analysis in social media. *Knowledge and Information Systems*, 60:617–663.

Huatian Zhang, Zhendong Mao, Kun Zhang, and Yongdong Zhang. 2022. Show your faith: Cross-modal confidence-aware network for image-text matching. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 3262–3270.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253.

Zhengxuan Zhang, Weixing Mai, Haoliang Xiong, Chuhan Wu, and Yun Xue. 2023. A token-wise graph-based framework for multimodal named entity recognition. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2153–2158. IEEE.

Weiyu Zhong, Zhengxuan Zhang, Qiaofeng Wu, Yun Xue, and Qianhua Cai. 2024. A semantic enhancement framework for multimodal sarcasm detection. *Mathematics*, 12(2):317.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 207–212, Berlin, Germany. Association for Computational Linguistics.

Yiyi Zhou, Tianhe Ren, Chaoyang Zhu, Xiaoshuai Sun, Jianzhuang Liu, Xinghao Ding, Mingliang Xu, and Rongrong Ji. 2021. Trar: Routing the attention spans in transformer for visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2074–2084.

Lei Zhu, Xinjiang Wang, Zhanghan Ke, Wayne Zhang, and Rynson WH Lau. 2023. Biformer: Vision transformer with bi-level routing attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10323–10333.

A Appendix

A.1 Dataset

We assess our model by conducting experiments on three publicly available benchmark datasets which are MVSA-Single, MVSA-Multiple (Niu et al., 2016), and HFM (Cai et al., 2019). MVSA-Single and MVSA-Multiple datasets collect data from Twitter, each text-image pair is labeled by a single sentiment. Both of them have three categories: positive, neutral, and negative. For a fair comparison, we process the original two MVSA datasets in the same way as Xu and Mao (2017). HFM dataset also collect data from Twitter, which has two sentimental categories: positive and negative. Following Cai et al. (2019), we adopt the same data preprocessing method for experiments. The statistics of these datasets are shown in Table 6.

A.2 Implementation Details

For a fair comparison, following the processing in Wei et al. (2023), we adopt the pre-trained BERT-base-uncased model (Kenton and Toutanova, 2019) as the text-encoder to embed each word of the text, and utilize the pre-trained ViT model (Dosovitskiy et al., 2020) as the image-encoder to embed each region of the image. The learning rate is $1e-5$ for MVSA-Single and MVSA-Multiple datasets, $2e-5$ for HFM dataset. We train the model for 20 epochs with mini-batch size 64. For MVSA-Single and MVSA-Multiple datasets, we establish the number of dynamic routing layers L as 4, the JS loss weight λ_1 as 0.9 and λ_2 as 0.3. For HFM dataset, we set the number of dynamic routing layer to 3, the JS loss weight λ_1 to 0.6 and λ_2 to 1.0. Adam optimizer is also utilized to train the model. Dropout and early stop are used to avoid overfitting. Based on prior configurations, we utilize ACC and Weighted F1 as evaluation metrics for the MVSA datasets and ACC and Macro-F1 for the HFM to assess the model's performance.

Table 6: Statistics of the dataset

Dataset	Training	Validating	Testing	Total
MVSA-S	3611	450	450	4511
MVSA-M	13624	1700	1700	17024
HFM	19816	2410	2409	24635

A.3 Baseline Models

We compare our model with unimodal baseline models and multimodal baseline models.

Unimodal Baselines. For text modality, we choose CNN (Kim, 2014), BiLSTM (Zhou et al., 2016), BERT (Kenton and Toutanova, 2019) and TGNN (Huang et al., 2019) as baselines since they are well-known models for text classification. For image modality, ResNet (He et al., 2016) and ViT (Dosovitskiy et al., 2020) are two popular models for image classification task, ODSA (Yang et al., 2020) is an image sentiment analysis model.

Multimodal Baselines. For MVSA-Single and MVSA-Multiple datasets, the baselines include: MultiSentiNet (Xu and Mao, 2017), a deep attention-based semantic network for multimodal sentiment analysis; HSAN (Xu, 2017), a hierarchical semantic attentional network based on image captions for multimodal sentiment analysis; Co-MN-Hop6 (Xu et al., 2018) utilize co-memory network to iteratively model the interactions between multiple modalities; MGNNS (Yang et al., 2021) adopt multi-channel graph neural networks with sentiment-awareness for image-text sentiment detection; CLMLF (Li et al., 2022) propose a contrastive learning and multi-layer fusion method for multimodal sentiment detection; MVCN (Wei et al., 2023) is the previous SOTA model that design a multi-view calibration network to solve the modality heterogeneity for multimodal sentiment detection. For HFM dataset, we compare two variants of Concat (Schifanella et al., 2016): Concat(2) means concatenating text and image, while Concat(3) introduces one more image attribute features; MMSD (Cai et al., 2019) is a hierarchical multimodal features model for fusing text, image, and image attributes; D&R Net (Xu et al., 2020) propose a decomposition and relation network to fuse the text, image, and visual attributes features.