Integrating Hierarchical Fine-Grained and Global Information for Multimodal Sentiment Analysis with Assistance of CLIP

Anonymous EMNLP submission

Abstract

Multimodal sentiment analysis aims to utilize the combined information from different modalities to gain a comprehensive understanding of human sentiment expressions. Previous research works have mostly focused on simple fusion of fine-grained features from text and images, neglecting the relationship between fine-grained features and high-level semantic global representations. In this paper, we propose a framework, dubbed INFIG, that integrates hierarchical fine-grained and global information to accurately capture sentiment expressions. We also leverage existing foundational models like CLIP to enhance the connection between similar image-text pairs and extract the latent high-level semantic information contained within weakly correlated image-text pairs. Extensive experiments on four publicly available multimodal datasets demonstrate the superiority and effectiveness of our proposed approach. The visualization further confirms the success of our model in integrating both fine-grained and global information, leading to better interpretability.

1 Introduction

Given the rising popularity of multimedia platforms and the advancement of information technology, there has been a significant surge in data that encompasses multiple modalities such as text, images, and video. Multimodal Sentiment Analysis (MSA) has gained considerable attention as a research area (Kaur and Kautish, 2022; Zhang et al., 2018; Yue et al., 2019). It plays a crucial role in domains like understanding human behavior, providing personalized services, and analyzing social media.

When dealing with multimodal data, the strong correlation between text and image can assist the model in effectively gauging the genuine sentiment behind the multimodal information, otherwise the opposite. In most cases, the weakly correlated modalities may contain more interference information. Such as Figure 1a, the girl's smile conveys pure joy in the snowy weather. Conversely, when analyzing Figure 1d with the accompanying text, it becomes evident that despite the girl's smile, the overall sentiment conveyed is one of sorrow and longing. If multiple modalities share similar emotional characteristics, there is an increased likelihood of making precise judgments regarding the polarity of the ultimate unified sentiment (Xu, 2017). However, if multiple modalities have different emotional characteristics, more complex judgment analysis will be required.

(a) Nothing beats the joy of waking up to a winter wonderland! #SnowyMornings #WinterBliss

(b) At the scene of a fatal collision on paramount near Upper Mount Albion Rd. @CHCHNews #hamont

(c) Happy spring! loving all the blossoming flowers happening here! so beautiful!

(d) So sad. 14th minute applause to commemorate a life gone far too soon

Figure 1: Examples of multimodal sentiment tweets

Some models (Xu, 2017; Kumar and Vepa, 2020) adopt a framework that involves training two encoders to extract visual and textual information. In this framework, the visual encoder typically relies on CNN architecture, while the text encoder is based on either RNN or Transformer models. Next, a fusion module is used to concatenate different modal features. (Xu and Mao, 2017) intro-

duces an LSTM model guided by visual features, to extract crucial words that determine the sentiment of the entire tweet and combined the representation of these words with visual semantic features, objects, and scenes. (Yang et al., 2020) proposes to use memory networks to realize the interaction between modalities. Some works develop a framework for multimodal multi-task learning based on late-fusion methods (Yu et al., 2020), or a network fusion model with residual connections based on late fusion (Ding et al., 2022). Despite the relative superiority of the above-mentioned models over unimodal models, the inputs with varying modalities are embedded in separate vector spaces. Hence, employing a simplistic concatenation approach without any pre-alignment operations to fuse the features of textual and visual data exhibits inferior performance.

Several early works have explored the implementation of contrastive learning techniques in the multimodal field. Such as CLIP (Radford et al., 2021) is trained using a contrastive loss on on a global similarity between its output embeddings. (Yuan et al., 2021) applies contrastive learning to learn visual representations in a unified multimodal training framework. In recent works, (Hu et al., 2022) fuses multimodal representation from multi-level textual information by injecting acoustic and visual signals into the T5 (Raffel et al., 2020) and acquires different multimodal representations by employing contrastive learning across modalities. (Gu et al., 2022) excels in leveraging fine-grained information between image encoder and text encoder. However, these works ignore the interplay between the interaction of fine-grained features and global features.

In this paper, we focus on visual-textual sentiment analysis in social media data and present an **IN**tegrating **FI**ne-grained and **G**lobal information(**INFIG**) method based on contrastive learning, which will help the model learn relevant or more profound levels of sentiment information within the textual and visual content.

Our contributions are summarized as follows:

1. We propose global alignment module and hierarchical fine-grained alignment module for mining more detailed semantic alignment. Furthermore, we investigate the optimal combination of these two modules, aiming to maximize their potential.
2. We use the pre-trained CLIP model to acquire implicit prior knowledge embedded within the visual and textual modalities, assisting global alignment module in capturing more precise multimodal features, thereby enhancing the model's capability to capture sentiment.
3. Our whole framework achieves better results than the state-of-the-art methods on most existing datasets, demonstrating its effectiveness and superiority.

2 Related Work

Multimodal Sentiment Analysis The objective of multimodal sentiment analysis is to extract emotions, interpretations, and sentiments by analyzing various modalities such as language, facial expressions, speech. (You et al., 2016) proposes a cross-modality consistent regression model to force the representations extracted from text and image to be consistent. (Zadeh et al., 2017) proposes a tensor fusion network that generates a novel tensor representation by performing the outer product operation on unimodal representations. (Zadeh et al., 2018) designs a memory fusion network for cross-view interactions. (Yu et al., 2021; Mai et al., 2020) focus on modal consistency and difference through multi-task joint learning and translating from one modality to another. (Ling et al., 2022) utilizes a combination of diverse pre-training tasks, such as masked language/region modeling and textual/visual opinion generation, to enhance the extraction of fine-grained aspect-based sentiment and the alignment across modalities. (Yang et al., 2021) applies a multi-channel graph neural network that is built based on the overall characteristics of the dataset. It incorporates a sentiment-awareness mechanism to perform multimodal sentiment analysis. (Zhu et al., 2022) proposes an innovative image-text interaction network that analyzes sentiments expressed in social media posts by leveraging the interaction between images and texts. (Liu et al., 2022) employs a hybrid curriculum learning approach to address the problem of semantic inconsistency between modalities. (Li et al., 2022) utilizes sentiment-label-based and data-augmentation-based contrastive learning to help the model capture the sentiment-related features in multimodal data.

Global and Fine-Grained Contrastive Learning Contrastive learning has gained major advances by viewing sample from multiple view, especially in representation learning. Its principle is quite clear,

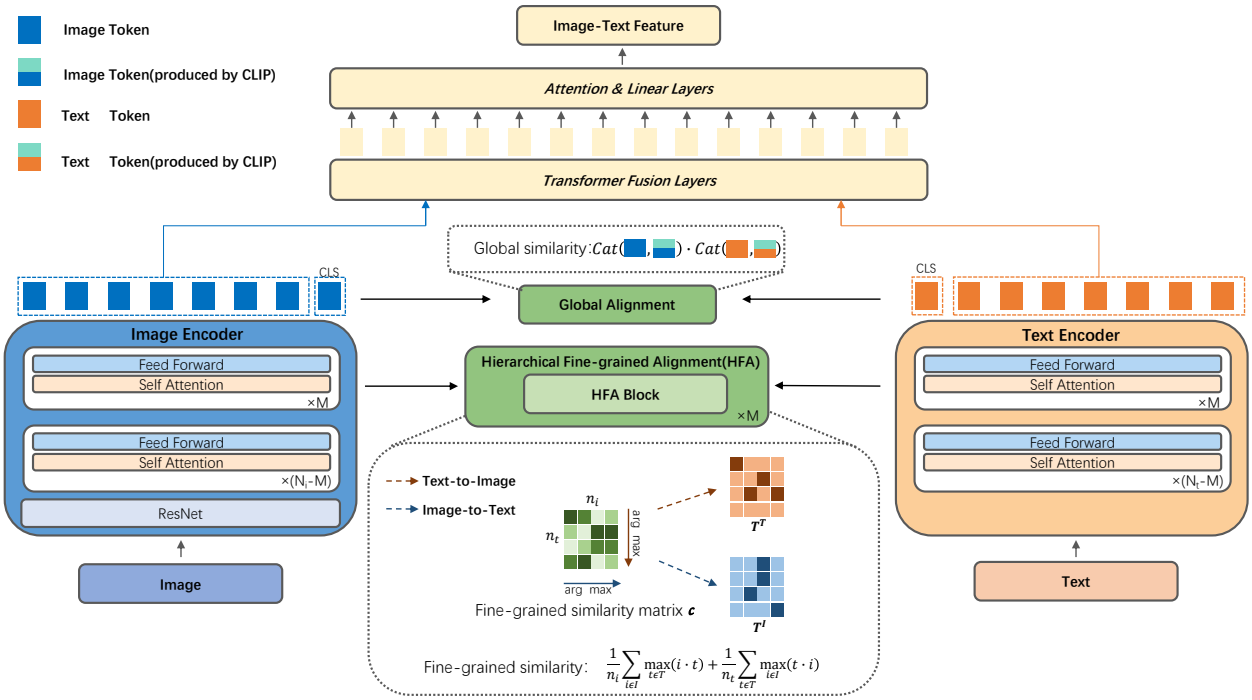


Figure 2: The overview of INFIG.

following the idea that an anchor and its positive sample should be pulled closer, while the anchor and negative samples should be pushed apart in feature space (Hadsell et al., 2006). (Gao et al., 2021; Yan et al., 2021) and (He et al., 2020; Chen et al., 2020; Chen and He, 2021) utilize global contrastive learning approach in the respective domains of natural language processing and computer vision. (Jia et al., 2021; Dou et al., 2022) encode visual and textual queries to global features and accordingly map them into a common latent space to compute the cosine similarity between two embedding vectors. Some efforts have been made to learn fine-grained cross-modal interaction between two modalities by leveraging token-wise or region-word similarities in the contrastive loss. (Lee et al., 2018) pays attention to fine-grained alignments by selectively attending to significant words or image regions. (Messina et al., 2021) detects and encodes image regions at the object level and sums the maximum of the region-word similarity scores with respect to each word or region. (Yao et al., 2021) aggregates the maximum token-wise similarity scores according to every single feature. Some works develop the reconstruction loss (Hazarika et al., 2020), or hierarchical mutual information maximization (Han et al., 2021) to achieve better modality fusion. Compared with the above works, we focus on the relationship between fine-grained features and global features, with the objective of effectively integrating them.

3 Method

3.1 Overall Framework

As shown in Figure 2, INFIG contains 5 components: an image encoder, a text encoder, a global alignment module, a hierarchical fine-grained alignment module and fusion layers.

We use ResNet (He et al., 2016) and N_i transformer layers as the image encoder to map the image into M+1 viusal token embeddings $\{i_1, ..., i_{n_i}, i_{cls}\}$ where i_{cls} is the embedding of the visual [CLS] token. We use debertav3 (He et al., 2021) based on N_t transformer layers as the text encoder to obtain the textual token representations $\{t_1, ..., t_{n_t}, t_{cls}\}$ where t_{cls} is the representation of the textual [CLS] token.

3.2 Global Alignment Module

Global Alignment Module aims to learn better representations at a global level before fusion. When the text and image are weakly correlated, meaning that the text may contain words unrelated to the image, it poses challenges to the fine-grained alignment learning in the model. In such cases, it

becomes crucial to enable the model to learn the deeper overall sentiment expressed by the image-text pair. We propose utilizing CLIP to provide a global prior information. Let i_{clip} and t_{clip} denote the representations of the image and text, respectively, obtained through linear projection after applying the CLIP model. Inspired by (He et al., 2020), we maintain two queues to store the most recent M image-text representations from the momentum unimodal encoders and adopt the following two InfoNCE (Oord et al., 2018) loss functions for global alignment module:

$$\begin{aligned} L_{i2t} &= -\log \frac{\exp(\sigma(i_g, t_g)/\tau)}{\sum_{m=1}^{M} \exp(\sigma(i_g, t_m)/\tau)} \\ L_{t2i} &= -\log \frac{\exp(\sigma(t_g, i_g)/\tau)}{\sum_{m=1}^{M} \exp(\sigma(t_g, i_m)/\tau)} \end{aligned} \tag{1}$$

where τ is a temperature factor, which is initialized as 0.07. We define $i_g = concat(i_{cls}, i_{clip})$, $t_g = concat(t_{cls}, t_{clip})$. Function σ computes the cosine similarity between two vectors. i_m and t_m are from momentum queues. The total loss of global alignment module is:

$$L_g = \frac{1}{2}(L_{t2i} + L_{i2t}) \tag{2}$$

3.3 Hierarchical Fine-Grained Alignment Module

The role of Hierarchical Fine-Grained Alignment Module is to find the most similarity textual token for each image patch, and similarly, for each textual token, to find its closest image patch. This enables the model to learn how the image and the text interact and influence each other, thereby enhancing its ability to accurately comprehend sentiment information. It takes $I = \{i_1, i_2, ..., i_{n_i}\}, I \in \mathbb{R}^{n_i \times d}$ and $T = \{t_1, t_2, ..., t_{n_t}\}, T \in \mathbb{R}^{n_t \times d}$ as inputs. For the k-th visual token i_k, we calculate its similarities with textual tokens and select the top j maximum similarity scores. Similarly, for textual tokens, we compute their similarity with visual tokens."

$$\begin{aligned} T_{i2t} &= Sum\{top_j(\sigma(i_k, t_1), ..., \sigma(i_k, t_{n_t}))\} \\ T_{t2i} &= Sum\{top_j(\sigma(t_k, i_1), ..., \sigma(t_k, i_{n_i}))\} \end{aligned} \tag{3}$$

In (3), we only consider the similarities between visual tokens and non-padding textual tokens. The similarities of p-th image to q-th text and the q-th text to the p-th image can be formulated as:

$$\begin{aligned} S_{p2q} &= \frac{1}{jn_i} \sum_{i \in I} T_{i2t} \\ S_{q2p} &= \frac{1}{jn_t} \sum_{t \in T} T_{t2i} \end{aligned} \tag{4}$$

and the fine-grained alignment loss of one layer can be expressed as follows:

$$\begin{aligned} L_I &= -\log \frac{\exp(S_{p2q}/\tau)}{\sum_{b=1}^{B} \exp(S_{p2q_b}/\tau)} \\ L_T &= -\log \frac{\exp(S_{q2p}/\tau)}{\sum_{b=1}^{B} \exp(S_{q2p_b}/\tau)} \end{aligned} \tag{5}$$

where B is the batch size. q_b and p_b are the rest of images and texts in batch excluding q and p. The total loss of fine-grained alignment module is:

$$L_f = \frac{1}{2M} \sum_{m=1}^{M} (L_T + L_I) \tag{6}$$

where m represents the layer m from the end. Our method is based on the recent FLIP approach, but instead of performing fine-grained alignment only at the last layer, we propose hierarchical fine-grained alignment at the last M layers of both text and image encoders,enabling the model to align deeper levels of information. And we select the top j similarity scores among tokens, rather than just the maximum, to enhance the robustness and generalization of model.

3.4 Fusion Layers

We use Transformer-Encoder as text-image fusion layers to fuse multimodal features. It is as follows:

$$\begin{aligned} F &= TE(i_1, ..., i_{n_i}, t_1, ..., t_{n_t}) \\ F &= \{f_1, f_2, f_3..., f_{n_i+n_t}\} \end{aligned} \tag{7}$$

where TE is the transformer-encoder. We obtain the sequence features F, but they cannot be directly used for the sentiment classification task. So we acquire the multimodal representation through the utilization of a straightforward attention layer.

$$\begin{aligned} G &= Attention(F) \\ G &= \{g_1, g_2, g_3..., g_{n_i+n_t}\} \end{aligned} \tag{8}$$

$$\tilde{K} = \sum_{i=1}^{n_t+n_i} g_i f_i \tag{9}$$

$$K = GELU(\tilde{K} W_K + b_K) \tag{10}$$

where G is the sequence of attention scores. GLUE is the activation function. $K \in \mathbb{R}^d$ is the mutimodal representation. Finally we use the cross-entropy loss as the sentiment classification loss:

$$L_{sc} = Cross - Entropy(GELU(KW + b)) \tag{11}$$

3.5 Model Training

The fine-grained alignment loss and global alignment loss can be incorporated into the total loss as regularization components.The total loss can be written as:

$$L = L_{sc} + \lambda_g L_g + \lambda_f L_f \tag{12}$$

where λ_g and λ_f are coefficients to balance the different training loss.

4 Experiments

4.1 Datasets

We conduct experiments on four publicly available visual-textual datasets which are MVSA-Single, MVSA-Multiple (Niu et al., 2016), HFM (Cai et al., 2019) and TumEmo (Yang et al., 2020). The detailed statistics of four datasets are shown in Table 1.

MVSA-Single and MVSA-Multiple are collected from Twitter posts. The former contains 4511 text-image pairs. Each pair is shown to a single annotator, who assigns one of three sentiments (positive, negative and neutral) to the text and image respectively. The latter contains 17024 image-text pairs, and each sample is labeled by three annotators. We use majority voting to obtain the single modality sentiment label. For a fair comparison, we process this two datasets in the same way uesd in (Li et al., 2022).

TumEmo is a large multimodal weak-supervision emotion dataset collected from Tumblr. Each text-image is categorized into one of seven sentiment classes (i.e., angry, bored, calm, fearful, happy, loving, and sad). We follow the same split of (Yang et al., 2021).

HFM is also collected from Twitter posts, which is contains 24635 image-text pairs. It is a binary sentiment dataset comprising positive and negative sentiments. We adopt the same data preprocessing method in (Cai et al., 2019).

4.2 Experimental Settings

We use debertav3-base as text encoder and ResNet-50, along with a 6-layer Transformer, as the image encoder. The number of the fusion layers is 3 for MVSA-Single, MVSA-Multiple, TumEmo and 4 for HFM. To save memory and scale up the batch size, automatic mixed-precision (Micikevicius et al., 2017) is used. The batch size is set to 64 for MVSA-Single, MVSA-Multiple, TumEmo and 128 for HFM. We use Adam optimizer. We set j is 2 for MVSA-Single, MVSA-Multiple and 3 for TumEmo and HFM. The learning rate is 2e-5. λ_f and λ_g are 0.2 and 0.3 respectively. The hierarchical fine-grained alignment module performs the last 3 Transformer layers of image encoder and text encoder, while the global contrastive is applied to the last layer. For more discussion on which layer global alignment module applies to, please refer to Section 4.6. All the experiments are done on four NVIDIA 4090 GPUS.

	Train	Valid	Test	All
MVSA-S	3611	450	450	4511
MVSA-M	13624	1700	1700	17024
HFM	19816	2410	2409	24635
TumEmo	156,204	19,525	19,536	195,265

Table 1: The detailed data splitting of MVSA-S, MVSA-M, TumEmo and HFM.

4.3 Baselines

We compare the proposed method with the following baselines:

MultiSentiNet (Xu and Mao, 2017) is a deep semantic network, which explicitly identifiers object and scene as semantic features of images with attention mechanism for multimodal sentimenti analysis.

Co-Memory (Xu et al., 2018) represents a co-memory network that employs an iterative approach to effectively model the interactions between multiple modalities.

MVAN (Yang et al., 2020) integrates Co-Memory and MultiSentiNet to enhance the modeling of correspondence between two modalities by learning scene-guided and object-guided text features as well as text-guided scene/object features.

MGNNS (Yang et al., 2021) introduces a multi-channel graph neural network that captures object, scene, and text representations by leveraging the global characteristics of the entire dataset.

Model	MVSA-Single		MVSA-Multiple		TumEmo		Model	HFM	
	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑		Acc↑	F1↑
MultiSentiNet	0.6984	0.6963	0.6886	0.6811	0.6418	0.5692	-	-	-
HSAN	0.6988	0.6690	0.6796	0.6776	0.6309	0.5398	Concat(2)	0.8103	0.7799
Co-Memory	0.7051	0.7001	0.6892	0.6883	0.6426	0.5909	Concat(3)	0.8174	0.7874
MVAN	0.7298	0.7139	0.7183	0.7038	0.6553	0.6543	MMSD	0.8344	0.8018
MGNNS	0.7377	0.7270	0.7249	0.6934	0.6672	0.6669	D&R Net	0.8402	0.8060
CLMLF	0.7533	0.7346	0.7200	0.6983	-	-	CLMLF	0.8543	0.8487
INFIG	**0.7641**	**0.7487**	0.7169	0.6972	**0.6781**	**0.6749**	INFIG	**0.8814**	**0.8780**

Table 2: Experimental results of different models on MVSA-Single, MVSA-Multiple, TumEmo and HFM datasets

CLMLF (Li et al., 2022) designs a label based contrastive learning and data based contrastive learning framework and fusion module to help the model learn general representations of image-text pairs.
(Schifanella et al., 2016) combines different features from multiple modalities to obtain multimodal representations. Concat(2) denotes the concatenation of textual features and image features, while Concat(3) includes an additional set of image attribute features.
MMSD (Cai et al., 2019) integrates text, visual and image attribute information using a hierarchical multimodal fusion model.
D&R Net (Li et al., 2022) constructs the Decomposition and Relation Network to fuse text, image and image attributes.

4.4 Results

Table 2 shows the quantitative comparison of our INFIG model with the baseline methods. We employ Weighted-F1 and ACC as the evaluation metrics for MVSA-Single, MVSA-Multiple and TumEmo, while Macro-F1 and ACC are used as the evaluation metrics for HFM. Compared to the previous SOTA, INFIG improves ACC of MVSA-Single, ACC of TumEmo, and ACC of HFM by 1.08%, 1.09%, and 2.71% respectively, and imporve F1 of MVSA-Single, F1 of TumEmo, and F1 of HFM by 1.41%, 0.8%, and 2.93% respectively. Due to the diverse types of multimodal information shared by users on social media, some instances contain consistent emotional information, while others are sparse and noisy. Our model still achieved the state-of-the-art performance in most cases. The experimental results presented above demonstrate that the proposed method of hierarchical fine-grained and global alignment can effectively leverage the implicit visual-textual information provided by CLIP to enhances the alignment of matched image-text pairs and explores deeper-level information within weakly correlated text-image pairs to learn the common features related to sentiment.

4.5 Ablation Study

We conducted a series of ablation studies on MVSA-Single, MVSA-Multiple, TumEmo and HFM datasets to evaluate the influence of Global Alignment Module(GAM), Hierarchical Fine-Grained Alignment Module(HFAM) and CLIP. The results are shown in 3. We can observe that removing the CLIP results in a decrease in both accuracy and F1 score, indicating the effectiveness of the implicit prior visual-textual information provided by the CLIP model. Additionally, we find that in certain cases, removing GAM results in a bigger performance degradation, while in other cases, removing HFAM has a more significant impact on performance decline. This indicates that GAM and HFAM exhibit varying degrees of importance in different situations. Moreover, the largest performance drop happens when GAM and HFAM are removed. It shows that the combination of GAM and HFAM produces better results than the individual utilization of either.

4.6 Influence of the GAM at Which Layer

We explore the optimal layer for applying the GAM. As shown in Figure 3, we fix HFAM in last three layers and conduct experiments using the last layer to the fourth-last layer of both the text encoder and the image encoder. We observe that when applying the GAM before the HFAM, specifically in the fourth-last layer, the performance is the poorest. The best performance is achieved when applying the GAM in the last layer. We attribute this to the possibility of the HFAM module disregarding the globaly information obtained by the GAM when placed before it. In contrast, by plcaing the GAM in the last layer, both the HFAM and GAM can jointly

Model	MVSA-Single		MVSA-Multiple		TumEmo		HFM	
	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑
INFIG	**0.7641**	**0.7487**	**0.7169**	**0.6972**	**0.6781**	**0.6749**	**0.8814**	**0.8780**
- w/o CLIP	0.7608	0.7423	0.7144	0.6946	0.6752	0.6698	0.8795	0.8729
- w/o GAM	0.7485	0.7336	0.7098	0.6813	0.6661	0.6630	0.8742	0.8699
- w/o HFAM	0.7433	0.7233	0.7051	0.6802	0.6679	0.6635	0.8760	0.8709
- w/o GAM, HFAM	0.7274	0.7193	0.6997	0.6738	0.6588	0.6561	0.8647	0.8600

Table 3: Ablation study of INFIG on MVSA-Single, MVSA-Multiple, TumEmo and HFM datasets

learn the global and fine-grained information of the multimodal features.

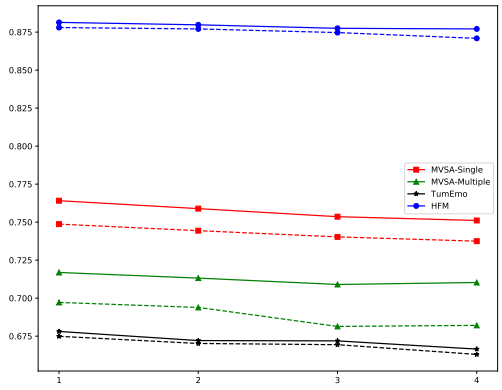
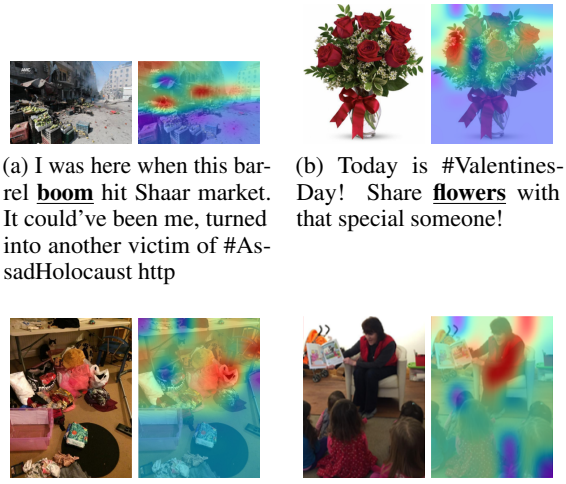


Figure 3: The solid line represents accuracy, the dashed line represents F1 score, and the x-axis indicates the layer at which the GAM module is applied from the last layer.

4.7 Visualization

To verify the effectiveness of GAM and HFAM, we visualize the attention weight of the Transformer-Encoder in the last layer of the fusion layers. Their visualization is shown in Figure 4. We can observe that the model can identify semantically related target on the image, even capturing target with deeper-level semantics, for a given keywords. This finding suggests that the model is capable of fine-grained alignment between the word in the text and the corresponding patch region in the image, as well as learning global information, which plays an important role in model for merging textual and visual features. For example, in Figure 4b, the model can effectively associate the word "flowers" with the target from the image. Furthermore, even though Figure 4a does not depict an actual boom, the model still manages to establish a correspondence between the aftermath of an explosion and the word "boom".



(a) I was here when this barrel **boom** hit Shaar market. It could've been me, turned into another victim of #AssadHolocaust http

(b) Today is #ValentinesDay! Share **flowers** with that special someone!

(c) houseguest coming to stay so I thought I'd tidy. It's turned into a **mass** sorting session.

(d) **Author** @CarolynjMorris reads to an enthusiastic audience #BigHeartDays @CreemoreOntario

Figure 4: Attention visualization of fine-grained and global alignment on multimodal datasets

5 Conclusion

We propose a simple and effective method for integrating hierarchical fine-grained and global information. Through visualization, the proposed method can be verified with intuitive interpretations. We also explore the optimal fusion strategy to achieve the best results and leverage the prior knowledge provided by CLIP to facilitate the learning of global information and intricate details of the visual-textual representations. Therefore, the model can capture the underlying sentiment expressions in the visual and textual content more accurately. We have achieved competitive performance compared to strong baseline models on public datasets.

Limitations

In this paper, we have only focused on alignment operations prior to the fusion layers. However, in the future, we plan to explore the utilization of more advanced techniques within the fusion layers to facilitate the learning of more correlated sentiment information across modalities. Additionally, we consider to incorporate the acoustic modality to enhance the model's capacity in acquiring richer sentiment features.

Ethics Statement

In this study, datasets are all open-source data for research purpose. Multimodal sentiment analysis has extensive applications in various domains, including social media analysis and intelligent robotics, and so on. However, achieving only 67.81% accuracy on a simple 7-class dataset TumEmo may not raise ethical or moral concerns in the real world. In practical applications, we will continue to monitor any potential ethical or moral issues.

References

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.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758.

Ning Ding, Sheng-wei Tian, and Long Yu. 2022. A multimodal fusion method for sarcasm detection based on late fusion. *Multimedia Tools and Applications*, 81(6):8597–8616.

Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. 2022. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Minzhe Niu, Hang Xu, Xiaodan Liang, Wei Zhang, Xin Jiang, and Chunjing Xu. 2022. Wukong: 100 million large-scale chinese cross-modal pre-training dataset and a foundation framework. *arXiv preprint arXiv:2202.06767*.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *arXiv preprint arXiv:2109.00412*.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1122–1131.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

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.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022. Unimse: Towards unified multimodal sentiment analysis and emotion recognition. *arXiv preprint arXiv:2211.11256*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

Ramandeep Kaur and Sandeep Kautish. 2022. Multimodal sentiment analysis: A survey and comparison. *Research Anthology on Implementing Sentiment Analysis Across Multiple Disciplines*, pages 1846–1870.

Ayush Kumar and Jithendra Vepa. 2020. Gated mechanism for attention based multi modal sentiment analysis. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4477–4481. IEEE.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pages 201–216.

Zhen Li, Bing Xu, Conghui Zhu, and Tiejun Zhao. 2022. Clmlf: a contrastive learning and multi-layer fusion method for multimodal sentiment detection. *arXiv preprint arXiv:2204.05515*.

Yan Ling, Rui Xia, et al. 2022. Vision-language pre-training for multimodal aspect-based sentiment analysis. *arXiv preprint arXiv:2204.07955*.

Huan Liu, Ke Li, Jianping Fan, Caixia Yan, Tao Qin, and Qinghua Zheng. 2022. Social image-text sentiment classification with cross-modal consistency and knowledge distillation. *IEEE Transactions on Affective Computing*.

Sijie Mai, Haifeng Hu, and Songlong Xing. 2020. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 164–172.

Nicola Messina, Giuseppe Amato, Andrea Esuli, Fabrizio Falchi, Claudio Gennaro, and Stéphane Marchand-Maillet. 2021. Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(4):1–23.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. 2017. Mixed precision training. *arXiv preprint arXiv:1710.03740*.

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.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

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.

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.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. 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 5065–5075, Online. Association for Computational Linguistics.

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.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2021. Filip: fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*.

Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2016. Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia. In *Proceedings of the Ninth ACM international conference on Web search and data mining*, pages 13–22.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3718–3727.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 10790–10797.

Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. 2021. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6995–7004.

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.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

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.

Tong Zhu, Leida Li, Jufeng Yang, Sicheng Zhao, Hantao Liu, and Jiansheng Qian. 2022. Multimodal sentiment analysis with image-text interaction network. *IEEE Transactions on Multimedia*.