ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446825

CLIP-MSA: INCORPORATING INTER-MODAL DYNAMICS AND COMMON KNOWLEDGE TO MULTIMODAL SENTIMENT ANALYSIS WITH CLIP

Qi Huang, Pingting Cai, Tanyue Nie, Jinshan Zeng⋆

School of Computer and Information Engineering, Jiangxi Normal University, Nanchang, China

{huangqi, cccai, tanyuenie, jinshanzeng}@jxnu.edu.cn

ABSTRACT

Multimodal Sentiment Analysis (MSA) aims to yield the sentiment polarities of speakers in video streams based on multiple modal features such as textual, acoustic and visual features, and has attracted amounts of attention in recent years. Existing MSA models often yield unimodal embeddings from the associated modal features individually, while overlooking the importance of inter-modal dynamics and common knowledge in the extraction of unimodal embeddings, resulting in the limited performance. In this paper, we suggest a novel MSA model called *CLIP-MSA* through incorporating the inter-modal dynamics and common knowledge into the generation of unimodal representations with the Contrastive Language-Image Pre-training (CLIP), and fusing the textual, acoustic and visual representations with a hierarchical co-attention mechanism. Numerous experimental results over two benchmark datasets show that the proposed model outperforms existing state-of-the-art models on CMU-MOSI, and provides competitive performance on CMU-MOSEI, in terms of four commonly used evaluation metrics.

***Index Terms*—** Multimodal sentiment analysis, CLIP, Hierarchical co-attention, Representation learning

1. INTRODUCTION

As an increasing number of users prefer to convey their viewpoints and emotions through videos, multimodal sentiment analysis (MSA) has attracted rising attention in recent years [1], due to its superiority on performance over the unimodal sentiment analysis, through utilizing multimodal features such as textual, acoustic and visual features to predict the sentiment intensities or polarities of speakers. The generation of unimodal representations to form the comprehensive multimodal representation plays a central role in MSA.

In MSA, previous works can be divided into two categories: methods that focus on representation learning and the methods that develop complex fusion mechanisms. As the typical methods of representation learning, MISA [2] exploited similarity and difference losses to obtain modality-invariant and -specific representations. HyCon [3] used hybrid contrastive learning to explore cross-modal interactions, learn inter-sample and inter-class relationships to generate high-quality unimodal representations. For the methods that focus on developing complex fusion mechanisms, Zadeh et al. [4] proposed a tensor fusion network, which used a three fold Cartesian product to fuse the information of different modalities. MulT [5] constructed unimodal and cross-modal Transformers to accomplish the fusion process through attention. Wang et al. [6] drew on translation models to explore subtle correlations between modalities and complete the fusion in the mutual conversion of source and target modality information. Although these methods achieve promising performance, they do not take the inter-modal dynamics and common knowledge into consideration in the extraction of unimodal embeddings, resulting in the limited performance.

Noticing that the inter-modal dynamics and common knowledge are important to enrich the unimodal representations, we incorporate them into unimodal representations with CLIP [7], motivated by the superiority of CLIP on capturing relationships at the semantic level and carrying a wealth of external knowledge. For each modality, we employ the associated encoder in CLIP to capture the inter-modal dynamics and common knowledge, and thus propose the CLIP enriched modal representation module. To effectively fuse the three modal representations enriched by CLIP, we suggest a hierarchical co-attention way according to the following fusion order, i.e., first fusing the visual and acoustic modal representations by a co-attention layer, and then the textual modal representation by another co-attention layer, inspired by the cognition mechanism of human-beings. The major contributions of this paper can be summarized as follows:

- We propose a novel MSA model called *CLIP-MSA* through incorporating inter-modal dynamics and common knowledge into the generation of unimodal representations with CLIP, and employing a hierarchical co-attention way to fuse the enriched multiple modal representations. By leveraging the pre-trained CLIP, the unimodal representations can be greatly enriched. The suggested hierarchical co-attention scheme aligns with the cognition mechanism of human-beings.
- A series of experiments are conducted over two benchmark datasets (CMU-MOSI and CMU-MOSEI) to show the effectiveness of the proposed model. Experimental results show that the proposed model is superior to state-of-the-art models on CMU-MOSI, and yields competitive results over CMU-MOSEI, in terms of four important evaluation metrics.

2. PROPOSED MODEL

In this section, we introduce the proposed CLIP-MSA model. An illustration of our model is given in Figure 1, which consists of three components: the CLIP enriched modal representation module, the hierarchical co-attention fusion module, and the sentiment predictor.

⋆Corresponding author

The work of J. Zeng was supported in part by the National Natural Science Foundation of China [Grant Nos. 62376110, 61977038], and Thousand Talents Plan of Jiangxi Province [Grant No. jxsq2019201124], and the Jiangxi Provincial Natural Science Foundation for Distinguished Young Scholars (20224ACB212004).

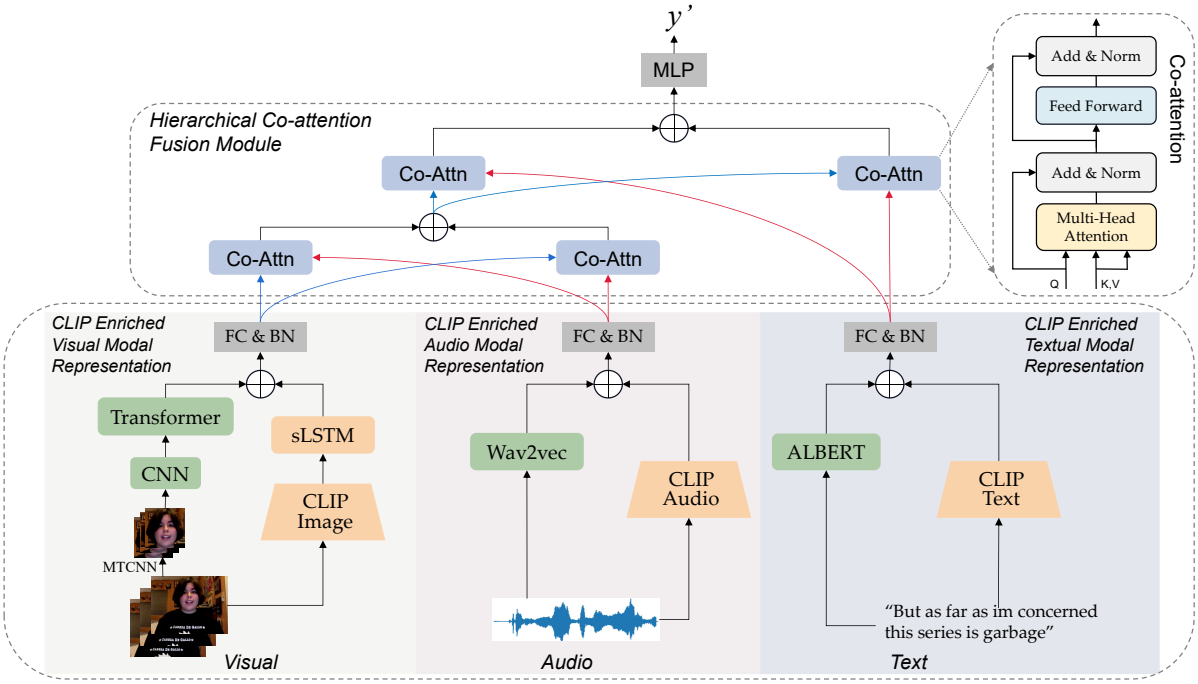

Fig. 1. Architecture of the proposed model CLIP-MSA.

2.1. Task Setup

Given a video U consisting of n utterances $U = [U_1, U_2, \ldots, U_n]$, where each utterance U_i contains a sequence I_t of text words, a sequence I_a of spectrogram chunks derived from the audio, and a sequence I_v of RGB image frames from the video, i.e., $U_i = (I_t, I_a, I_v)_i$. The goal of the MSA task is to construct a function $f : U_i \rightarrow y_i'$ that maps each utterance U_i to the corresponding sentiment label y_i'.

2.2. CLIP Enriched Modal Representations

For each modality, we employ a dual-branch strategy for feature extraction. One branch utilizes encoders specific to the three modalities to extract distinct information. CLIP is used as another branch to explore richer unimodal information from a cross-modal perspective.

2.2.1. Textual Modal Representation

The pre-trained ALBERT [8] is employed as one branch for text feature extraction. We input the raw text sequence I_t to ALBERT, and select the vector at the [CLS] position from the final layer as the representation for the sentence, denoted as F_t^{albert}. The another branch utilizes the text encoder from the pre-trained CLIP model to obtain the CLIP-encoded text feature F_t^{clip}. Finally, the utterance-level textual representation $F_t \in \mathbb{R}^d$ is obtained by concatenating the outputs from the two branches and subjecting the concatenated representation to a linear and a BatchNorm [9] transformation.

2.2.2. Audio Modal Representation

For the audio modality, one branch employs the pre-trained wav2vec 2.0 [10], which has shown exceptional performance in the audio domain. We calculate its utterance-level feature $F_a^{wav2vec}$ by taking the mean along the first dimension. The other branch utilizes the pre-trained ESResNeXt [11] model, which is jointly trained with CLIP using contrastive learning. The final discourse-level acoustic representation $F_a \in \mathbb{R}^d$ is also obtained by concatenating the features from the two branches above and then feed it to a linear and a BatchNorm transformation.

2.2.3. Visual Modal Representation

Similar to much of the prior research, we consider facial expressions to be a crucial channel for conveying sentiments within the visual modality. Therefore, for one branch of visual feature extraction, we initially employ the MTCNN [12] to detect the facial position within image frames. Subsequently, this information is fed into a pre-trained CNN model (an 11-layer VGG model). And then utilize the Transformer [13] to encode the embedding of image sequence, yielding F_v^{trans}. In another branch, we start by using the image encoder from the pre-trained CLIP model to encode the original image frames. Then we utilize a sLSTM [14] to model the temporal information, obtaining the CLIP-encoded visual feature F_v^{clip}. The ultimate discourse-level visual representation $F_v \in \mathbb{R}^d$ is also obtained by concatenating the features from the two branches and then passing through a linear and a BatchNorm transformation.

2.3. Hierarchical Co-attention Fusion

To effectively fuse the features of three modalities, we devised a hierarchical fusion approach based on co-attention. Each layer of fusion consists of two co-attention (CT) units. As illustrated in Figure 1, a CT unit comprises a multi-head attention network, a feed-forward neural network, and two sets of residual connections and layer normalization applied subsequently.

2.3.1. Co-attention Unit

In CT, two distinct modality inputs are defined as X_1 and X_2. X_1 serves as the query Q, while X_2 is used for the key K and value V. The CT calculates the co-attention matrix for each head as follows:

$$h_i = Softmax(\frac{(X_1 W_i^Q)(X_2 W_i^K)^T}{\sqrt{d_h}}) X_2 W_i^V$$

where the parameter matrices $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{1 \times d_h}$, $d_h = d/m$ is the dimensionality of the output feature of each head, while m is the number of heads.

The multi-head attention is obtained by concatenating all the co-attention matrices, i.e., $h = (h_1; h_2; ...; h_m)W^O$, where $W^O \in \mathbb{R}^{md_h \times 1}$. Afterward, h and X_1 pass through a two-layer normalized FFN to obtain attention-based multimodal representations:

$$F_{1 \leftarrow 2} = Norm(Norm(X_1 + hX_1) + FFN(Norm(X_1 + hX_1)))$$

where $F_{1 \leftarrow 2}$ is a modality 2 weighted 1 feature. Meanwhile, in another CT unit of the same layer, X_2 is used as the query Q, while X_1 is employed for the key K and value V, performing the aforementioned operations and yields a modality 1 weighted 2 feature $F_{2 \leftarrow 1}$. Finally, we concatenate $F_{1 \leftarrow 2}$ and $F_{2 \leftarrow 1}$ and passing them through a linear layer yields the multimodal fusion features for this stage:

$$F_{12} = Linear([F_{1 \leftarrow 2}; F_{2 \leftarrow 1}], \theta_{12}^{linear})$$

where θ_{12}^{linear} is the parameter of linear layer.

2.3.2. Hierarchical Fusion

The text contains high-level semantic information, requiring time and context for understanding. Therefore, in the hierarchical fusion process, we first combine visual and audio features, followed by the fusion of the combined audio-visual representation with text.

In the first stage of fusion, we employ F_a and F_v as the inputs. Passing through two CT units, we can obtain the audio-visual multimodal representation F_{av}.

In the second stage, we use F_{av} and F_t as inputs to the next two CT units to obtain the fusion representation F_m, which incorporates information from all three modalities.

2.4. Sentiment Inference

The multimodal representation F_m obtained from the hierarchical fusion process is fed into a fully connected deep neural network to generate the final sentiment scores y'. In this paper, we utilize the SmoothL1Loss [15] to compute the loss between the predicted value y' and the ground truth y, defined as $L_{task} = \frac{1}{N} \sum_{i=1}^{N} SmoothL1Loss(y_i, y_i')$, where N is the number of samples.

3. EXPERIMENTS

3.1. Datasets and Metrics

CMU-MOSI is one of the most widely used benchmark datasets in the field of MSA. It consists of 93 movie review videos collected from YouTube, segmented into 2199 short monologue video clips at the utterance level. Each video clip is annotated with sentiment intensity ranging from -3 (strongly negative) to 3 (strongly positive). **CMU-MOSEI** is an extended version of CMU-MOSI. It comprises 23,453 video segments from 3,228 videos, involving up to 1,000 speakers and covering 250 topics. Similar to MOSI, each utterance-level sample in MOSEI is annotated a sentiment label on a scale of -3 to 3.

Following the prior work [16], we use four metrics to evaluate models, which are binary classification accuracy (Acc-2), F1-Score, Mean Absolute Error (MAE), and Pearson Correlation Coefficient. Notably, Acc-2 and F1-Score are calculated using two distinct schemes: negative/non-negative, which includes 0, and positive/negative, which excludes 0.

3.2. Baselines

To examine the performance of CLIP-MSA, we compared our model with numerous prominent methods, including methods based on learning: TFN [4], LMF [17] and MFM [18]; Works based on fusion: MulT [5] and MAG-BERT [19]; The methods based on feature space: MISA [2]; Multi-task-based works: Self-MM [16] and TPMSA [20]; And the methods based on contrastive learning: MMIM [21], HyCon [3], and WSCL-CL [22].

3.3. Implement Details

Similar to [11], we choose the ResNet-based CLIP model for the visual and textual modalities and a pre-trained ESResNeXt model that is jointly trained with CLIP for the audio modality. We conduct three separate experiments using different random seeds and present the averaged performance. All experiments were conducted using an Nvidia 3080Ti GPU.

3.4. Experimental Results

3.4.1. Quantitative Results

Table 1 presents a comparison between our model and the baseline models in terms of results. As shown in Table 1, our approach outperforms the baseline models across all evaluation metrics in CMU-MOSI. In CMU-MOSEI, our model surpasses all methods in Acc-2 (has0) and F1-score (has0), and demonstrates comparable performance in other metrics as well. Furthermore, we faithfully reproduced two excellent baseline models, Self-MM [16] and MAG-BERT [19], as outlined in their original papers. Our model surpasses them across various metrics. These results indicate that compared to methods like MISA [2], Self-MM [16] and MIMM [21], which rely on single-branch feature extraction through loss-controlled learning, our proposed dual-branch unimodal feature extraction method, enriched with CLIP, demonstrates greater advancement. When compared to various fusion strategies such as tensor-based fusion TFN [4], cross-modal attention-based fusion MulT [5] and TPMSA [20], our hierarchical co-attention fusion mechanism also achieves superior performance.

3.4.2. Ablation Study

To further distinguish the contribution of each part in CLIP-MSA, we conducted a series of ablation experiments on the CMU-MOSI dataset, with results shown in Table 2.

To verify the effectiveness of the CLIP feature extraction branch and hierarchical co-attention fusion mechanism, we conducted ablation experiments, as shown in the first part of Table 2. When removing the CLIP-based feature extraction branch (w/o CLIP) for all three modalities, Acc-2 and F1-Score decrease by 2.37/2.65 and 2.42/2.66, respectively. When eliminating the hierarchical fusion mechanism based on co-attention (w/o Co-Attn, replacing it

Models	MOSI				MOSEI			
	Acc-2	F1-Score	MAE↓	Corr	Acc-2	F1-Score	MAE↓	Corr
TFN(2017'EMNLP)†	-/80.8	-/80.7	0.901	0.698	-/82.5	-/82.1	0.593	0.700
LMF(2018'ACL)†	-/82.5	-/82.4	0.917	0.695	-/82.0	-/82.1	0.623	0.677
MFM(2019'ICLR)†	-/81.7	-/81.6	0.877	0.706	-/84.4	-/84.3	0.568	0.717
MulT(2019'ACL)†	81.50/84.10	80.60/83.90	0.861	0.711	-/82.50	-/82.30	0.580	0.703
MISA(2020'ACM MM)†	81.80/83.40	81.70/83.60	0.783	0.761	83.60/85.50	83.80/85.30	0.555	0.756
MAG-BERT(2020'ACL)†	82.54/84.30	82.59/84.30	0.731	0.789	83.79/85.23	83.74/85.08	0.539	0.753
Self-MM(2021'AAAI)†	84.00/85.98	84.42/85.95	0.713	0.798	82.81/85.17	82.53/85.30	0.530	0.765
MMIM(2021'EMNLP)‡	84.14/86.06	84.00/85.98	0.700	0.800	82.24/85.97	82.66/85.94	0.526	0.772
HyCon(2022'IEEE)‡	-/85.50	-/85.40	0.688	0.818	-/86.40	-/86.40	0.590	0.788
TPMSA(2022'IEEE)‡	-/87.0	-/87.0	0.704	0.799	-/85.6	-/85.6	0.542	0.770
WSCL-CL(2022'EMNLP)‡	-/86.3	-/86.2	0.712	0.798	-/86.1	-/86.0	0.577	0.794
Self-MM*	83.27/85.15	83.19/85.14	0.712	0.795	81.95/84.94	82.34/84.87	0.532	0.762
MAG-BERT*	82.92/84.58	82.96/84.58	0.770	0.770	81.43/85.22	80.95/85.18	0.556	0.755
CLIP-MSA(ours)	**85.32/87.30**	**85.39/87.31**	**0.673**	**0.820**	**84.19**/85.19	**84.16**/85.43	0.563	0.761

Table 1. Results on CMU-MOSI and CMU-MOSEI. The best results are marked in bold. † is from [16] and ‡ is from the corresponding original papers. ∗: reproduced from open-source codes with hyper-parameters provided in original papers. In Acc-2 and F1-Score, the left of the "/" is calculated as "negative/non-negative" and the right is calculated as "negative/positive"

Model	Acc-2	F1-Score	MAE↓	Corr
CLIP-MSA	**85.32/87.30**	**85.39/87.31**	**0.673**	**0.820**
w/o CLIP	82.95/84.65	82.97/84.65	0.728	0.795
w/o Co-Attn	83.28/85.11	83.27/85.05	0.721	0.805
w/o CLIP, Co-Attn	82.31/84.50	82.36/84.48	0.782	0.769
w/o CLIP-T	84.01/85.72	84.02/85.68	0.711	0.804
w/o CLIP-A	83.62/85.47	83.66/85.46	0.732	0.790
w/o CLIP-V	83.43/85.31	83.49/85.31	0.725	0.800
(A,T,V)	84.31/86.69	84.40/86.72	0.737	0.786
(V,T,A)	84.65/86.49	84.69/86.48	0.717	0.798

Table 2. Ablation studies of CLIP-MSA over CMU-MOSI dataset.

with concatenation), Acc-2 and F1-Score decrease by 2.04/2.19 and 2.12/2.26, respectively. When the above two are removed (w/o CLIP, Co-Attn), all metrics experience further decline. These results demonstrate that the introduced CLIP feature extraction branch is able to extract important information from the data to form high-quality unimodal representations, and suggest that the hierarchical co-attention fusion method effectively integrates information from different modalities to create comprehensive multimodal representations.

Meanwhile, we conducted ablation experiments to investigate the impact of CLIP features on different modalities. As shown in the second part of Table 2, we individually removed the CLIP extraction branch from the textual (w/o CLIP-T), audio (w/o CLIP-A), and visual (w/o CLIP-V) modal feature extraction processes. It can be observed that the performance experienced a decrease of 1 to 2 points in all cases. Notably, the impact of removing the CLIP branch from the text modality is relatively smaller compared to the other two modalities. We attribute this to the excellent performance of pre-trained language models. Visual and audio features often contain a large amount of noise, requiring additional information for supplementation and noise removal. Therefore, when removing the CLIP branch from the audio and visual modalities, a more significant decrease in performance is observed.

Furthermore, we conducted ablation experiments on the fusion order, as depicted in the third part of Table 2. '(A,T,V)' denotes the sequence of fusing audio and text information before visual information, while '(V,T,A)' signifies the order of fusing visual and text information prior to audio information. It can be observed that both '(A,T,V)' and '(V,T,A)' exhibit a decline in performance compared to CLIP-MSA's fusion order that integrates acoustic and visual features before textual feature. This is because text often contains a higher semantic space than visual and audio modalities. Notably, '(V,T,A)' performs better than '(A,T,V)' in certain metrics. We attribute this to the fact that CLIP is trained on a large amount of image-text pairing data, while ESResNXet is jointly trained with CLIP on a general dataset. Consequently, the features encoded by CLIP's text and image encoders are more aligned in the feature space, resulting in a slightly more effective fusion of textual and visual information compared to text and audio fusion.

4. CONCLUSION

In this paper, we explore the performance of CLIP in MSA. We propose a CLIP-MSA model, introducing the CLIP encoders as the second branch for unimodal feature extraction, which utilizes the extensive external knowledge carried by CLIP and its ability to capture advanced semantic information to generate high-quality unimodal representations. In addition, we designed a hierarchical co-attention fusion method to integrate information from different modalities in a suitable order. Extensive experiments show the superior performance of CLIP-MSA. In our future work, we will further excavate the application of CLIP in MSA, exploring the role of CLIP in aligning modality features.

5. REFERENCES

[1] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea, "Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research," *IEEE Transactions on Affective Computing*, 2020.

[2] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 1122–1131.

[3] Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu, "Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis," *IEEE Transactions on Affective Computing*, 2022.

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

[5] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*. NIH Public Access, 2019, vol. 2019, p. 6558.

[6] Zilong Wang, Zhaohong Wan, and Xiaojun Wan, "Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis," in *Proceedings of The Web Conference 2020*, 2020, pp. 2514–2520.

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

[8] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

[9] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. pmlr, 2015, pp. 448–456.

[10] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.

[11] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel, "Audioclip: Extending clip to image, text and audio," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 976–980.

[12] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE signal processing letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[14] Alex Graves and Alex Graves, "Long short-term memory," *Supervised sequence labelling with recurrent neural networks*, pp. 37–45, 2012.

[15] Ross Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[16] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu, "Learning modality-specific representations with self-supervised multitask learning for multimodal sentiment analysis," in *Proceedings of the AAAI conference on artificial intelligence*, 2021, vol. 35, pp. 10790–10797.

[17] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency, "Efficient low-rank multimodal fusion with modality-specific factors," *arXiv preprint arXiv:1806.00064*, 2018.

[18] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov, "Learning factorized multimodal representations," *arXiv preprint arXiv:1806.06176*, 2018.

[19] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque, "Integrating multimodal information in large pretrained transformers," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*. NIH Public Access, 2020, vol. 2020, p. 2359.

[20] Bo Yang, Lijun Wu, Jinhua Zhu, Bo Shao, Xiaola Lin, and Tie-Yan Liu, "Multimodal sentiment analysis with two-phase multi-task learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2015–2024, 2022.

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

[22] Sijie Mai, Ya Sun, and Haifeng Hu, "Curriculum learning meets weakly supervised modality correlation learning," *arXiv preprint arXiv:2212.07619*, 2022.