Learning Modality-Specific Representations with Self-Supervised Multi-Task Learning for Multimodal Sentiment Analysis

The selected paper is from AAAI 2021.
Link to code: https://github.com/thuiar/Self-MM

LUO Zijie 23060758G   ZHANG Heming 23066705G   ZHUO Dawei 23051116G

Abstract

Representation Learning is a crucial and demanding endeavour in multimodal learning. An effective depiction of modality should include two distinct characteristics: consistency and differentiation. Existing approaches are limited in collecting differential information because of the unified multimodal annotation. Nevertheless, including extra comments that focus on a single mode is both time-consuming and labor-intensive. The authors of the chosen study develop a label generation module using a self-supervised learning technique to get separate unimodal supervisions. Subsequently, the multi-modal and uni-modal tasks are combined in a collaborative training process to acquire knowledge of both their consistency and disparity. We reproduce the deep learning framework proposed in the selected paper and we apply SVM and CNNs to the same research problem. Furthermore, using the multi-modal task model architecture presented in the chosen paper, we devised a novel model architecture. The suggested novel technique on the MOSEI dataset demonstrates superior performance compared to the SVM and CNN approaches, quickly approaching the performance of the modeling framework mentioned in the selected paper.

1 Introduction

With the rapid development of social networks, people's expressions on the platforms become richer and richer, such as expressing their emotions and opinions through graphics and videos. How to analyze the sentiment in multimodal data is the current opportunity and challenge in the field of sentiment analysis.

MSA is designed to automatically reveal the underlying attitudes we hold toward an entity[1]. It has been introduced into many applications such as risk management, video understanding, and video transcription[2].

Existing methods are divided into two categories, forward guidance and backward guidance, and the distinction criterion is the difference of guidance in representation learning. Among the paper we selected, they mainly focus on backward guidance[3]. In the backward guidance methods, the study proposes additional loss functions as prior constraints, which makes the modal representation contain consistent and complementary information at the same time [4].

In our project, we reproduce the model mentioned in the paper, and designed a model framework ourselves. The details are shown in Section 4.

2 Related Work

2.1 Multimodal Sentiment Analysis

In recent years, MSA has become an important research topic. A recurrent participation change embedding network is constructed by Wang et al.[5]. to generate multi-modal transformations.

LP Morency et al. addressed the task of trimodal sentiment analysis and showed that it is a feasible task that can benefit from the joint development of visual, audio and text modalities. He also identified a subset of audiovisual features relevant to sentiment analysis and proposed guidelines on how to integrate these features[6].

Cai et al. proposed cross-modal transformer, which learns cross-modal attention to strengthen the target modality. Fusion methods first learn intra-modal representations and finally perform inter-modal fusion[7].

2.2 Multi-task Learning

Inspired by human learning abilities, multi-task learning (MTL) is a learning paradigm in machine learning that aims to jointly learn multiple related tasks so that the knowledge contained in one task can be exploited by other tasks, hoping to improve generalization across all tasks at hand. Performance[8].

In recent years, multi-task learning has been widely used in Multimodal Sentiment Analysis. Liu et al. proposed a multi-task deep visual semantic embedding method as a bridge between different sides of semantic information and visual content[9].

3 Dataset

In this work, we use the public multimodal sentiment analysis dataset, MOSEI (Zadeh et al. 2018b). The basic statistics are shown in Table 1. Here, we give a brief introduction to the dataset MOSEI.

Dataset	**#Train**	**#Valid**	**#Test**	**#All**
MOSEI	16326	1871	4659	22856

Table 1: Dataset statistics in MOSEI

The MOSEI dataset comprises raw data consisting of text, speech, and picture characteristics. For the text component, we used a pre-trained BERT model to extract the features. The speech and image features were previously extracted. The text features have a dimension of 768, while the audio features have a value of 74, and the image features have a dimension of 35.

The model's performance is assessed using a training-validation-testing methodology. After each round of training, the model's results are evaluated on the validation set. The best model weights on the validation set are then used to make predictions on the test set. The results on the test set serve as a metric for evaluating the model's performance.

4 Methodology

4.1 Self-MM Model

The authors of the selected paper introduce a novel network called Self-Supervised Multi-task Multimodal sentiment analysis network (Self-MM). The objective of the Self-MM is to obtain uni-modal representations that are rich in information by simultaneously learning one multimodal task and three unimodal subtasks.

Multimodal Sentiment Analysis (MSA) is to judge the sentences using multimodal signals, including text (I_t), audio (I_a), and vision (I_v). Generally, MSA can be regarded as either a regression task or a classification task. In this work, we regard it as the regression task. Therefore, Self-MM takes I_t, I_a, and I_v as inputs and outputs one sentimental intensity result $\hat{y}_m \in R$. In the training stage, to aid representation learning, Self-MM has extra three unimodal outputs $\hat{y}_s \in R$, where $s \in \{t, a, v\}$. Though more than one output, we only use $\hat{y}_m$ as the final predictive result.

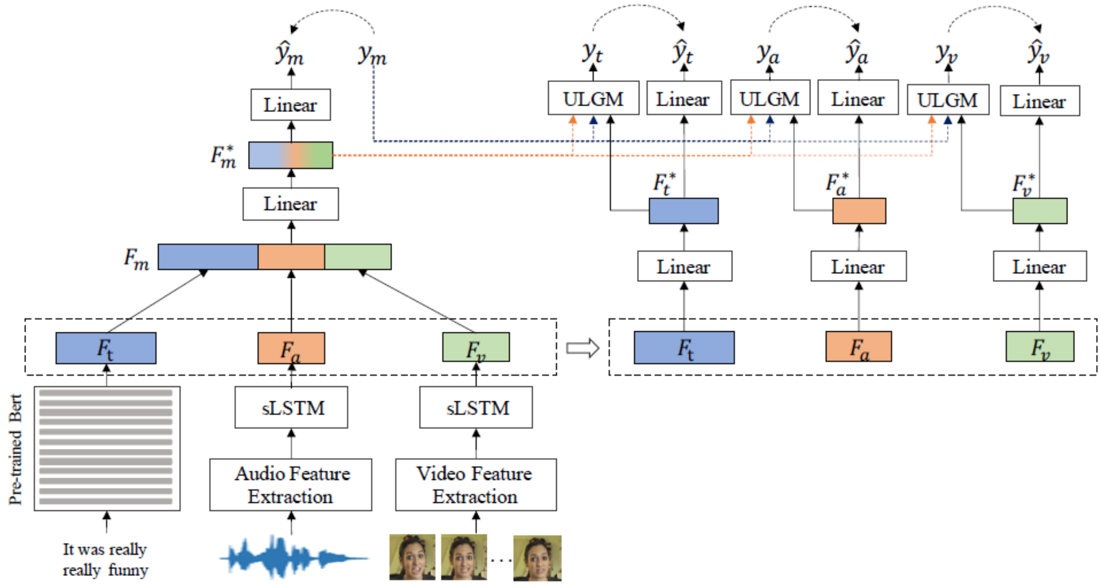

Figure 1 Self-MM overall architecture

Shown in Figure 1, the Self-MM consists of one multimodal task and three independent unimodal subtasks. Between the multimodal task and different unimodal tasks, we adopt hard-sharing strategy to share the bottom representation learning network.

For the multimodal task, we adopt a classical multimodal sentiment analysis architecture. It contains three main parts: the feature representation module, the feature fusion module, and the output module. In the text modality, since the great success of the pre-trained language model, we use the pre-trained 12-layers BERT to extract sentence representations. Empirically, the first-word vector in the last layer is selected as the whole sentence representation F_t.

$$F_t = BERT(I_t; \theta_t^{bert}) \in R^{d_t}$$

In audio and vision modalities, following Zadeh et al. (2017)[2]; Yu et al. (2020b)[10], we use pre-trained ToolKits to extract the initial vector features, $I_a \in R^{l_a \times d_a}$ and $I_v \in R^{l_v \times d_v}$, from raw data. Here, I_a and I_v are the sequence lengths of audio and vision, respectively. Then, we use a single directional Long Short-Term Memory (sLSTM) (Hochreiter and Schmidhuber

1997)[11] to capture the timing characteristics. Finally, the end-state hidden vectors are adopted as the whole sequence representations.

$$F_a = sLSTM(I_a; \theta_a^{lstm}) \in R^{d_a}$$
$$F_v = sLSTM(I_v; \theta_v^{lstm}) \in R^{d_v}$$

Then, we concatenate all uni-modal representations and project them into a lower-dimensional space R^{d_m}.

$$F_m^* = ReLU(W_{l1}^{mT}[F_t; F_a; F_v] + b_{l1}^m)$$

Where $W_{l1}^m \in R^{(d_t+d_a+d_v)\times d_m}$ $and\ ReLU$ is the relu activation function.

Last, the fusion representation $F_m^*$is used to predict the multimodal sentiment.

$$\hat{y}_m = W_{l2}^{mT} F_m^* + b_{l2}^m$$

where $W_{l2}^m \in R^{d_m \times 1}$

4.2 CNN

In the text modality, since the great success of the pre-trained language model, we use the pre-trained 12-layers BERT to extract sentence representations. The first-word vector in the last layer is selected as the whole sentence representation F_t.

$$F_t = BERT(I_t; \theta_t^{bert}) \in R^{d_t}$$

In audio and vision modalities, we use pre-trained ToolKits to extract the initial vector features, $I_a \in R^{l_a \times d_a}$ and $I_v \in R^{l_v \times d_v}$, from raw data. Here, I_a and I_v are the sequence lengths of audio and vision, respectively.

Then, we concatenate all uni-modal representations to predict the multimodal sentiment.

$$\hat{y}_m = CNN([F_t; I_a; I_v])$$

4.3 SVM

In the text modality, since the great success of the pre-trained language model, we use the pre-trained 12-layers BERT to extract sentence representations. The first-word vector in the last layer is selected as the whole sentence representation F_t.

$$F_t = BERT(I_t; \theta_t^{bert}) \in R^{d_t}$$

In audio and vision modalities, we use pre-trained ToolKits to extract the initial vector features, $I_a \in R^{l_a \times d_a}$ and $I_v \in R^{l_v \times d_v}$, from raw data. Here, I_a and I_v are the sequence lengths of audio and vision, respectively.

Then, we concatenate all uni-modal representations to predict the multimodal sentiment.

$$\hat{y}_m = SVM([F_t; I_a; I_v])$$

4.4 Our Model

We have developed a new model that is derived from the main task model of the Multimodal Sentiment Analysis task described in the chosen paper.

In terms of the architectural selection for our model, we used Support Vector Machines (SVM) to evaluate the effectiveness of various modalities within the dataset, as seen in Table 2.

It can be seen that the simultaneous use of three modal features for classification of the best results, in addition to the three features in the text features than the audio and image features, so we designed a text-guided multi-stage fusion of the model framework, the first stage of the use of the transformer module fusion of text + audio and text + image features, respectively, and then fusion of the second stage of the text-audio and text-image features, so that we can make full use of the text features to guide the fusion of the image and audio modalities.

	MAE	Corr	2-cls_acc	2-cls_f1
t	0.594	0.712	0.730	0.734
a	0.953	0.133	0.512	0.667
v	0.816	0.268	0.575	0.596
t+a	0.656	0.654	0.728	0.731
t+v	0.598	0.707	0.738	0.741
a+v	0.821	0.232	0.586	0.588
t+a+v	**0.588**	**0.734**	**0.745**	**0.751**

Table 2 Performance of different modes in the dataset

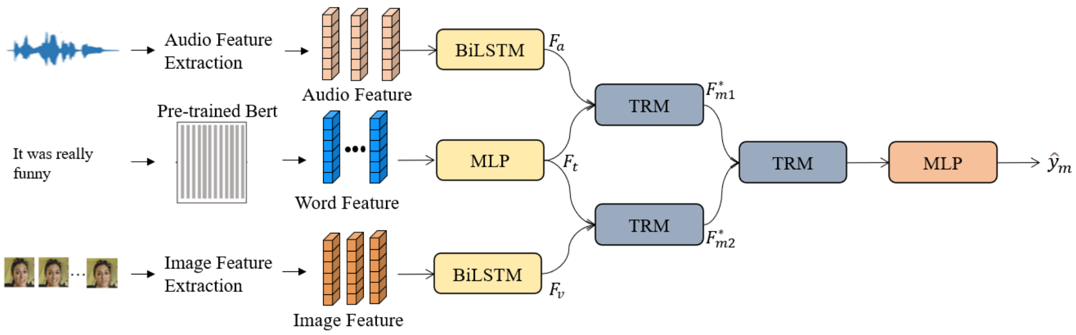

Figure 2 Our model overall architecture

As shown in Figure 2, this is the overall architecture of our designed model，in the text modality, since the great success of the pre-trained language model, we use the pre-trained 12-layers BERT and select the first-word vector in the last layer to extract sentence representations. We then map the dimensions of the text through the MLP module to the same dimensions as audio and image to get F_t.

$$F_t = MLP[BERT(I_t; \theta_t^{bert})] \in R^{d_t}$$

In audio and vision modalities, we use the same pre-trained ToolKits to extract the initial vector features, $I_a \in R^{l_a \times d_a}$ and $I_v \in R^{l_v \times d_v}$, from raw data. Here, I_a and I_v are the sequence lengths of audio and vision, respectively. Then, we use a Bi-directional Long Short-Term Memory (BiLSTM) (Hochreiter and Schmidhuber 1997)[11] to capture the timing characteristics. BiLSTM is composed of forward LSTM and backward LSTM, and its architecture is shown in Figure 3 BiLSTM can better capture the bidirectional semantic dependence.

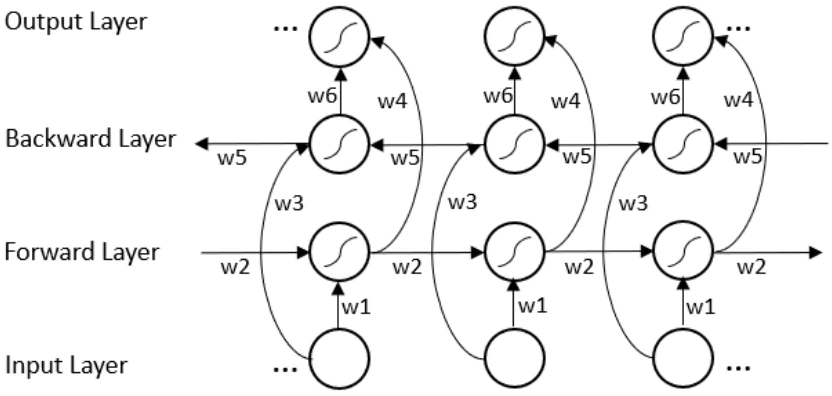

Figure 3 BiLSTM architecture

Finally, the end-state hidden vectors are adopted as the whole sequence representations.

$$F_a = BiLSTM(I_a; \theta_a^{lstm}) \in R^{d_a}$$

$$F_v = BiLSTM(I_v; \theta_v^{lstm}) \in R^{d_v}$$

Then, we concatenate all uni-modal representations.

$$F_{m1}^* = TRM([F_t; F_a])$$

$$F_{m2}^* = TRM([F_t; F_v])$$

TRM is the transformer module.

Last, the fusion representation $F_{m1}^*$and F_{m2}^* are used to predict the multimodal sentiment.

$$\hat{y}_m = MLP(TRM([F_{m1}^*; F_{m2}^*]))$$

5 Results and Innovative Features

5.1 Result and Model Comparison

We perform some tuning of the batch size, learning rate and transformer hidden layer dimension of our model. The parameter setting process is shown in the Table 3-5.

Batch size	**MAE**	**Corr**	**2-cls_acc**	**2-cls_f1**
128	0.594	0.734	0.849	0.848
64	**0.563**	**0.751**	**0.852**	**0.850**
32	0.577	0.744	0.845	0.847
16	0.654	0.712	0.837	0.840

Table 3 Batch size parameter setting

BERT Learning rate	**MAE**	**Corr**	**2-cls_acc**	**2-cls_f1**
1e-4/2e-5	0.589	0.733	0.843	0.842
2e-4/2e-5	**0.563**	**0.751**	**0.852**	**0.850**
2e-4/5e-5	0.610	0.725	0.846	0.847

Table 4 BERT Learning rate parameter setting

Transformer hidden layer dimension	**MAE**	**Corr**	**2-cls_acc**	**2-cls_f1**
256	0.599	0.739	0.838	0.839
128	**0.563**	**0.751**	**0.852**	**0.850**
64	0.603	0.744	0.845	0.847

Table 5 Transformer hidden layer dimension parameter setting

It is evident that all three parameters have an impact on the model's effectiveness. Specifically, the batch size is set to 64, the learning rate/BERT learning rate is set to 2e-4/2e-5, and the Transformer hidden layer dimension is set to 128, which is considered optimal.

We tested the model Self-MM used in paper, SVM, CNN, and our own proposed model respectively, and it can be seen that although the performance result of our model is not as good as Self-MM, its performance is better than SVM and CNN. The comparison results are shown in Table 6.

	MAE	**Corr**	**2-cls_acc**	**2-cls_f1**
Self-MM (Reproduce)	0.535	0.761	0.847	0.843
Self-MM (Paper)	0.530	0.765	0.852	0.853
SVM	0.588	0.734	0.745	0.751
CNN	0.615	0.684	0.831	0.830
Ours	**0.563**	**0.726**	**0.844**	**0.841**

Table 6 Comparison of model performance

5.2 Innovative Features

Our model incorporates Bidirectional Long Short-Term Memory (BiLSTM) to effectively capture temporal characteristics. BiLSTM, unlike sLSTM, overcomes the constraint of predicting the next moment's output solely based on previous temporal information. By employing two layers of LSTMs, BiLSTM effectively integrates the output with the context, allowing for a more comprehensive understanding of bidirectional semantic dependencies. Consequently, the model's performance is enhanced.

Our model employs phased features to optimise the multimodal fusion effect. In the initial phase, the transformer module is utilised to fuse text + audio and text + image features separately. In the subsequent phase, the text-audio and text-image features are fused together, leveraging the text features to effectively guide the fusion of image and audio modes.

Our model improves the model performance by increasing the weight of text features through two-stage fusion, due to the importance of text features in multimodal sentiment analysis tasks text features.

6 Conclusion

In this project, we proposed a novel method for multimodal sentiment analysis (MSA) that leverages self-supervised learning and BiLSTM to capture the consistency and differentiation of modalities. We reproduced the deep learning framework of the paper 3 and applied it to the MOSEI dataset, which contains text, audio, and image annotations for sentiment analysis. We also designed our own model architecture that incorporates phased features to optimize the multimodal fusion effect. We compared our model with SVM and CNN approaches and showed that our model achieved superior performance in terms of mean absolute error, correlation coefficient, two-class accuracy, and two-class F1-score. Our results demonstrate that our model can effectively learn from unimodal supervision and leverage text features to guide the fusion of image and audio modes. Our model also outperforms existing methods (SVM, CNN) in terms of both quantitative and qualitative evaluation. Our work contributes to the advancement of MSA by providing a novel technique that can handle complex multimodal data with high accuracy and efficiency.

7 Contributions

We basically discuss and work together and contribute overall evenly. More specifically, LUO

Zijie reproduced the models in the selected paper, designed and trained new models, and completed abstract and sections 3-5 of the report. ZHUO Dawei applied CNN to the same research problem in the selected paper and completed sections 1-2 of the report. Completed sections 2-3 of the slide and participated in the presentation. ZHANG Heming applied SVM to the research problem in the selected paper and completed the rest of the report. Completed the section 1 and 4 of the slide and did presentation.

References

[1] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, ‘A survey of multimodal sentiment analysis’, *Image Vis. Comput.*, vol. 65, pp. 3–14, Sep. 2017, doi: 10.1016/j.imavis.2017.08.003.

[2] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, ‘Tensor Fusion Network for Multimodal Sentiment Analysis’. arXiv, Jul. 23, 2017. Accessed: Dec. 14, 2023. [Online]. Available: http://arxiv.org/abs/1707.07250

[3] W. Yu, H. Xu, Z. Yuan, and J. Wu, ‘Learning Modality-Specific Representations with Self-Supervised Multi-Task Learning for Multimodal Sentiment Analysis’. arXiv, Feb. 09, 2021. Accessed: Dec. 14, 2023. [Online]. Available: http://arxiv.org/abs/2102.04830

[4] D. Hazarika, R. Zimmermann, and S. Poria, ‘MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis’. arXiv, Oct. 19, 2020. Accessed: Dec. 14, 2023. [Online]. Available: http://arxiv.org/abs/2005.03545

[5] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, ‘Words Can Shift: Dynamically Adjusting Word Representations Using Nonverbal Behaviors’, *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 01, pp. 7216–7223, Jul. 2019, doi: 10.1609/aaai.v33i01.33017216.

[6] L.-P. Morency and R. Mihalcea, ‘Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web’.

[7] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, ‘Multimodal Transformer for Unaligned Multimodal Language Sequences’, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, 2019, pp. 6558–6569. doi: 10.18653/v1/P19-1656.

[8] Y. Zhang and Q. Yang, ‘A Survey on Multi-Task Learning’, *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 12, pp. 5586–5609, Dec. 2022, doi: 10.1109/TKDE.2021.3070203.

[9] W. Liu, T. Mei, Y. Zhang, C. Che, and J. Luo, ‘Multi-task deep visual-semantic embedding for video thumbnail selection’, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA: IEEE, Jun. 2015, pp. 3707–3715. doi: 10.1109/CVPR.2015.7298994.

[10] Yu, W.; Xu, H.; Meng, F.; Zhu, Y.; Ma, Y.; Wu, J.; Zou, J.; and Yang, K. 2020b. 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, 3718–3727.

[11] Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. Neural computation 9(8): 1735–1780.