UTILIZING BERT INTERMEDIATE LAYERS FOR MULTIMODAL SENTIMENT ANALYSIS

Wenwen Zou, Jundi Ding, Chao Wang*

School of Computer Science and Engineering, Nanjing University of Science and Technology
zww2020@njust.edu.cn, 1596864655@qq.com, Superwang@njust.edu.cn

ABSTRACT

Some recent works use pre-trained BERT to extract text features instead of the GloVe embedding representation, which greatly improves multimodal sentiment analysis. However, these works ignore BERT's intermediate layers information. The layers in BERT can capture phrase-level, syntax-level, and semantic-level information, respectively. Utilizing these levels of information in the multimodal fusion stage can lead to fine-grained fusion results and promote the potential of fine-tuning BERT on multimodal data. In this paper, we fuse middle layers information of BERT with non-verbal modalities in multiple stages via our designed hierarchical fusion structure external to BERT. In addition, the crossmodal fusion process runs the risk of discarding valid information of unimodality. We suggest distilling sentiment-relevant features from the removed information and restitute it to the network to promote sentiment analysis. Evaluating our proposed model on CMU-MOSI and CMU-MOSEI datasets, we show that it outperforms existing works and successfully fine-tunes BERT on multimodal language data.

Index Terms— multimodal sentiment analysis, BERT, fine-tune, feature restitution

1. INTRODUCTION

When we communicate with people in daily life, we convey our attitudes and emotions by utilizing a mix of language, acoustic, and vision modalities. Multimodal Sentiment Analysis (MSA) is an increasingly growing research area that aims at analyzing the sentiment of multimodal communication. A considerable challenge in this area is to combine three modal information effectively.

Previous work mainly focus on designing sophisticated multimodal fusion strategies to explore intramodal and intermodal dynamics. Tensor-based networks such as Tensor Fusion Network (TFN) [1] and Low-rank Multimodal Fusion Network (LMF) [2] adopt outer product to capture unimodal, bimodal, trimodal interactions. Attention-based networks exploit various attention mechanism components to fuse modalities: Recurrent Attended Variation Embedding Network (RAVEN) [3] learns multimodal-shifted word representations conditioned on the visual and acoustic modalities through the attention gating mechanism; Multimodal Transformer (MulT) [4] merges unaligned multimodal time-series via multiple directional pairwise crossmodal transformers which translates source modality into the target using cross-attention.

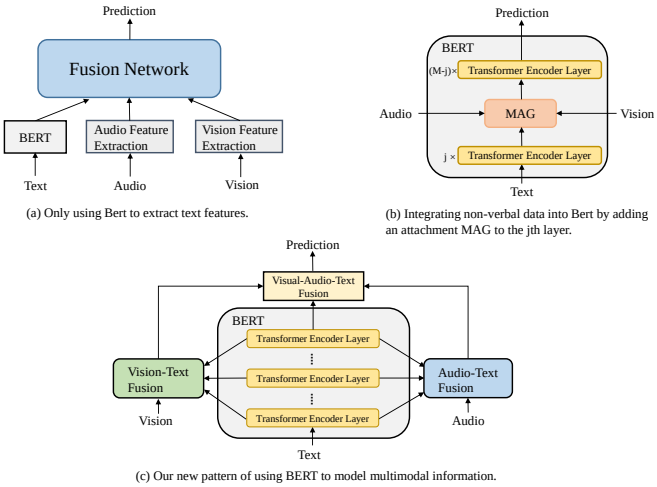


Fig. 1. Architecture examples of works using BERT for multimodal tasks.

Works in [5] investigated which aspects of these fusion methods could effectively solve multimodal language analysis problems. They observed that text is the most informative modality, and using it as a pivot for visual and acoustic modalities achieved improved performance. More recently, some studies have tried to improve multimodal sentiment analysis by using excellent text feature representation. In 2018, the pre-trained model BERT (Bidirectional Encoder Representations from Transformers) [6] was born. Using BERT to extract text features can get an excellent contextual representation, making it a great success in many NLP tasks. However, the exploration of applying BERT to multimodal datasets is still limited. There are currently two main ways (see Fig. 1): (a) Hazarika et al. use BERT to extract text features, and then use their designed network [7] to process multimodal features; (b) Rahman et al. integrate non-verbal information

*Corresponding author.

2022 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-8563-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICME52920.2022.9860014

on the one of layers of BERT by a middleware called Multimodal Adaptive Gate (MAG) [8] and implement downstream tasks through fine-tuning. However, the work of (a) do not fine-tune BERT on multimodal data, and they only use the final output representation of BERT, ignoring the rich semantic information in the middle layers. The work of (b) makes BERT seamlessly adapt to multimodal input by using a simple component. Nevertheless, according to the description of the paper, non-verbal data only directly affect the features of one layer in the BERT. With the deepening of the internal modeling process in BERT, from shallow to high layers can learn phrase-level, syntactic-level, and semantic-level information [9]. If non-verbal modalities interact with text features at different stages, it can obtain fine-grained fusion features.

Based on the above observation, we propose an effective framework that can fine-tune BERT on multimodal language data. We design a multimodal hierarchical fusion structure external to BERT, where acoustic and visual features will be used as supplementary information to fuse with multiple intermediate layers information of BERT hierarchically. In order to enhance the memory ability of the hierarchical fusion framework, we introduce Gated Recurrent Unit (GRU) [10] to memorize the fusion features of each stage. Moreover, the process of multimodal fusion can potentially lose some discriminative sentiment features in unimodality. Given the significance of text modality, we attach importance to restoring sentiment-relative features of text modality to the network. Inspired by [11], we intend to use a dual restitution loss to disentangle the discarded information into sentiment-relevant and sentiment-irrelevant features. This loss ensures that after restoring the sentiment-relevant information to the network, the features will become more discriminative, promoting sentiment analysis.

To summarize, our main contributions are as follows:

- We propose a new way of using BERT to model multimodal information, which makes full use of different intermediate layers of BERT, realizes hierarchical multimodal fusion, and successfully fine-tunes BERT on multimodal language data.
- We introduce a restitution module, which can distill sentiment-relative feature from removed information using the dual restitution loss and restore it to the network to promote sentiment analysis.
- Our experiments on CMU-MOSI and CMU-MOSEI benchmark datasets demonstrate the effectiveness of our multimodal sentiment analysis system.

2. METHODOLOGY

2.1. Problem formulations

The input to the modal is an utterance, a segment of a video bounded by pauses and breaths. For each utterance U, the input comprises of three modalities, audio (a), visual (v) and text (t). These are represented as $U_t \in \mathbb{R}^{T\times d_t}$, $U_v \in \mathbb{R}^{T\times d_v}$ and $U_a \in \mathbb{R}^{T\times d_a}$, where T represents the sequence length of unimodality, and d_t, d_v and d_a denote feature dimensions of the text, vision and audio modality, respectively. The details of these features are discussed in Section 3.2.

Given these sequences $U_{m\in\{t,a,v\}}$, the primary task is to predict sentiment polarity (negative/positive) of utterance U by scoring a continues intensity variable $y \in \mathbb{R}$.

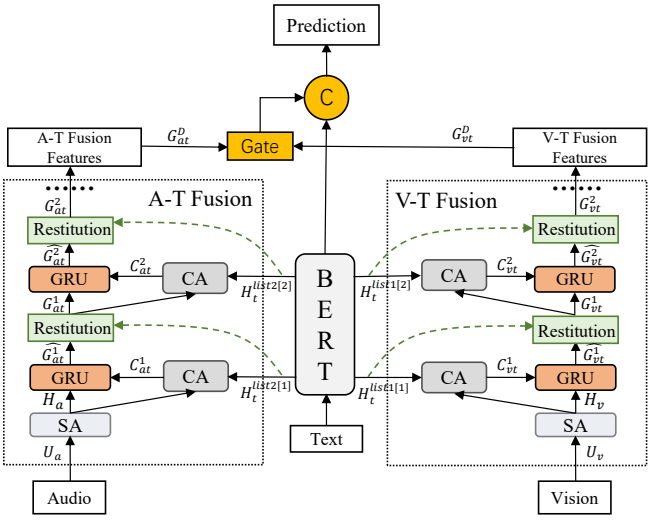


Fig. 2. Overall architecture for MHMF-BERT.

2.2. Proposed model

We present an overview of Multimodal Hierarchical Memory Fusion network using BERT intermediate layers (MHMF-BERT) in Fig.2. The architecture of our system can be mainly divided into three parts: unimodal representation, bimodal fusion, and trimodal fusion.

2.2.1. Unimodal representation

To obtain the unimodal latent space of the three modalities, we leverage BERT to process the input text, and use H_t^i to represent the i^{th} intermediate layer information of BERT. For visual and audio modalities, we use Self-attention Adapter (SA) module to modal their intramodal dynamics and denote H_v, H_a as module output. SA is composed of two transformer encoder layers [12] (see Fig.3(b)), the self-attention mechanism in the encoder can capture the long-term dependencies of the sequence.

$$H_t^i = BERT^{i^{th}\ layer}(U_t) \tag{1}$$

$$H_m = SA(U_m), \quad m \in \{v, a\} \tag{2}$$

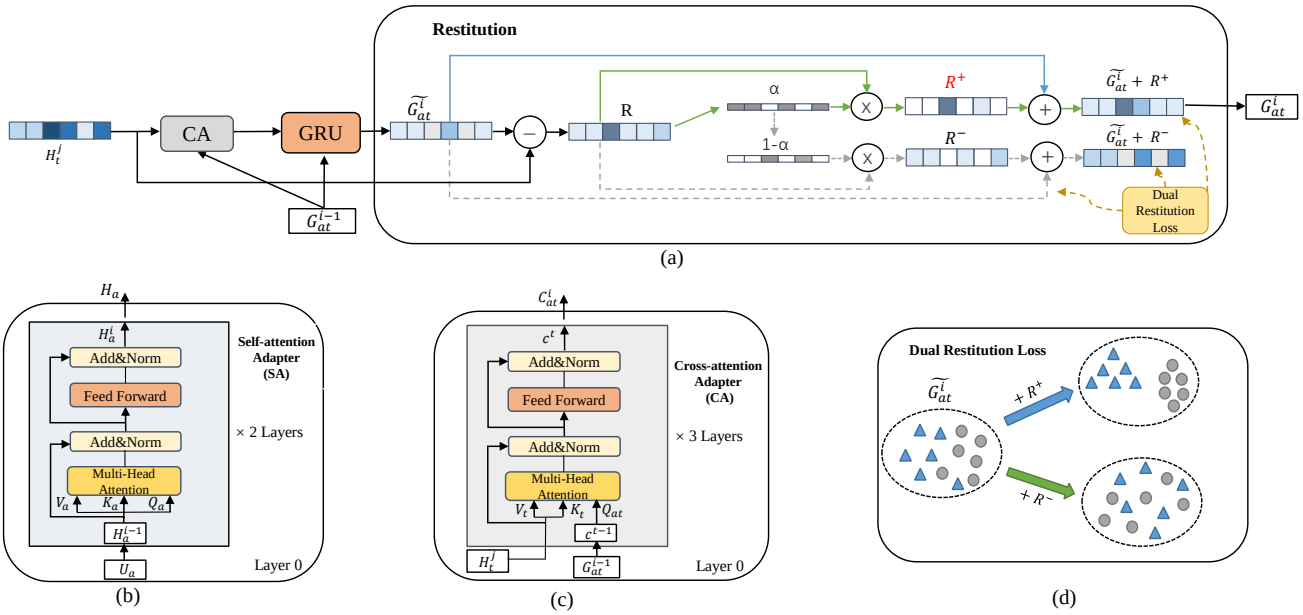


Fig. 3. (a) Proposed Restitution module. The crossmodal fusion operation runs the risk of removing some sentiment-relevant features of unimodality, after which we perform valuable feature reconstruction (marked with solid green arrows). (b) A Self-attention Adapter is deep stacking of 2 Transformer encoder layers. (c) A Cross-attention Adapter comprises three crossmodal attention layers that enable one modality to receive information from another modality. (d) Dual restitution loss encourages the disentanglement of the removed features R into task-relevant features R^+ and task-irrelevant features R^-, when adding them to network, they will increase and decrease discrimination, respectively.

2.2.2. *Bimodal fusion*

Crossmodal attention mechanism. We cross-fuse the visual/audio modality with intermediate layer information of BERT. The fusion part is mainly completed by multiple Cross-attention Adapter(CA) modules, and each CA comprises three crossmodal-attention layers [4] (see Fig.3(c)). CA serves to repeatedly reinforce a target modality with features from another source modality by learning the attention across the two modalities' features. The cross-fusion operation for a pair of audio-text modalities can be expressed as:

$$C_{at}^{i} = CA_{t\to a}(H_t^{list2[i]},\ G_{at}^{i-1}), \quad i \in [1, D] \tag{3}$$

Where $list2[\cdot]$ is a list parameter, which controls the selection of layer in BERT for fusing with audio modality features (e.g., list2[1]=2 means selecting the second layer information of BERT H_t^2), and G_{at}^{i-1} denotes the $i-1^{th}$ Restitution module output, we elaborate on the module hereafter.

Gate recurrent memory. Our model architecture is a hierarchical network, and there is a risk of forgetting information. The GRU has fewer parameters and is simpler than other recurrent models. We adopt GRU to keep bimodal fusion information to enhance the memory capability of our network.

$$G_{at}^{0} = H_a \tag{4}$$

$$\widetilde{G_{at}^{i}} = GRU(C_{at}^{i},\ G_{at}^{i-1}), \quad i \in [1, D] \tag{5}$$

Restitution. The process of fusing multimodal information can potentially remove sentiment-relevant features of unimodality. Given the importance of text modality, we perform valuable feature reconstruction. Firstly, we propose to disentangle the discarded information R into sentiment-relevant feature R^+ and sentiment-irrelevant feature R^- by a learned attention vector a.

$$R = H_t^{list1[i]} - \widetilde{G_{at}^{i}}, \quad i \in [1, D] \tag{6}$$

$$a = \sigma(W_2\delta(W_1 Avgpool(R))) \tag{7}$$

$$R^+ = aR \tag{8}$$

$$R^- = (1-a)R \tag{9}$$

where W_1 and W_2 are two FC layers, $\sigma(\cdot)$ denotes sigmoid activation function, $\delta(\cdot)$ denotes ReLU activation function.

Then adding the sentiment-relevant feature R^+ to $\widetilde{G_{at}^{i}}$, we obtain the output feature G_{at}^{i} of the i^{th} Restituiton module as

$$G_{at}^{i} = \widetilde{G_{at}^{i}} + R^+ \tag{10}$$

2.2.3. *Trimodal fusion*

In order to emphasize the useful features and overcome noise, we use gated vector to fuse vision-text fusion vector and audio-text fusion vector.

$$g_{vt} = ReLU(W_{vt} \cdot [G_{vt}^{D},\ H_v] + b_{vt}) \tag{11}$$

$$g_{at} = ReLU(W_{at} \cdot [G^D_{at},\, H_a] + b_{at}) \tag{12}$$

$$F_{vat} = g_{vt} \cdot G^D_{vt} + g_{at} \cdot G^D_{at} \tag{13}$$

2.2.4. Prediction

Finally, we concatenate the gated fusion feature with BERT's last layer output and then feed it to the prediction layer: pool them and pass an affine transformation to generate the value that can predict the label.

2.3. Model learning

The overall learning of the model is performed by minimizing:

$$L = L_{task} + \sum_{i=1}^{D} \lambda_i L^i_{restitution} \tag{14}$$

Here, λ_i is the interaction weight that determine the contribution of corresponding regularization component to the overall loss L. $L^i_{restitution}$ denotes the restitution loss for the i^{th} Restitution module. Each of these component losses are responsible for achieving the desired subspace properties. We discuss them next.

2.3.1. Restitution Loss

As illustrated in Fig. 3(d), the main idea of restitution loss is that: after adding the sentiment-relevant features R^+ to fused vector, the enhanced features become more discriminative; on the other hand, after adding the sentiment-irrelevant features R^- to fused vector, the features become less discriminative. We facilitate the disentanglement of residual features by comparing the discrimination capability of features before and after the restitution.

For the simplicity of the formula, here we express the restituting effective feature operation in each Restitution module as $F^+ = F + R^+$. And $F^- = F + R^-$ denotes adding sentiment-irrelevant feature to original feature.

We randomly drawn three samples from each mini-batch, i.e., an anchor sample a, a positive sample p which has the same sentiment label with the anchor, a negative sample n that has a different label from the anchor. We differentiate the three samples by subscript: anchor sample (F_a, F^+_a, F^-_a), positive sample (F_p, F^+_p, F^-_p), negative sample (F_n, F^+_n, F^-_n)

The restitution loss is thus defined as:

$$\begin{aligned} L^+_{restitution} = {} & Softplus(d(F^+_a, F^+_p) - d(F_a, F_p)) \\ & + Softplus(d(F_a, F_n) - d(F^+_a, F^+_n)) \end{aligned} \tag{15}$$

$$\begin{aligned} L^-_{restitution} = {} & Softplus(d(F_a, F_p) - d(F^-_a, F^-_p)) \\ & + Softplus(d(F^-_a, F^-_n) - d(F_a, F_n)) \end{aligned} \tag{16}$$

$$L_{restitution} = L^+_{restitution} + L^-_{restitution} \tag{17}$$

where $d(x, y)$ denotes the distance between x and y, which is defined as $d(x, y) = 0.5 - x^T y/(2||x||\,||y||)$. $Softplus(\cdot) = ln(1 + exp(\cdot))$ is a monotonically increasing function that reduces optimization difficulty by avoiding negative loss values.

2.3.2. Task Loss

The task-specific loss estimates the quality of prediction during training. We use mean squared error loss to train our task. For N_b utterances in a batch, these are calculated as:

$$L_{task} = \frac{1}{N_b} \sum_{i=0}^{N_b} ||y_i - \hat{y}_i||^2_2 \tag{18}$$

3. EXPERIMENT

3.1. Experimental Settings

We evaluate our proposed approach on two high-quality multimodal data, CMU-MOSI [13] and CMU-MOSEI [14]. For both MOSI and MOSEI, each sample is labeled by human annotators with a sentiment score in the range [-3, +3], -3 is strongly negative, and +3 means strongly positive.

According to the previous works, we use the following metrics to evaluate the performance of the model: 1) Acc2: binary accuracy, positive or negative; 2) F1 score; 3) MAE: mean absolute error; 4) Corr: the correlation between the model's prediction and that of humans. Except for MAE, higher values denote better performance for all metrics.

3.2. Baselines

To comprehensively compare our method, we list several baselines and state-of-the-art models for MSA.

TFN (Tensor Fusion Network) [1] calculates a multi-dimensional tensor which capture unimodal, bimodal, trimodal interactions across linguistic, visual and acoustic modalities to model intra-modality and inter-modality dynamics explicitly.

LMF (Low-rank Multimodal Fusion) [2] is an upgraded version of TFN, it applies low-rank multimodal tensors fusion technique to reduce the number of parameters.

MulT (Multimodal Transformer) [4] merges multimodal time-series via multiple directional pairwise crossmodal transformers. Each crossmodal transformer is a deep stacking of several crossmodal attention blocks, and it can implicitly match different modal flows.

ICCN (Interaction Canonical Correlation Network) [15] first fuses audio and video modality features with text embeddings to get two outer products, text-audio, and text-video. Then, the outer products are fed to a Canonical Correlation Analysis (CCA) network, whose output is used for prediction.

MISA (Modality-Invariant and -Specific Representations) [7] incorporate a combination of losses including distributional similarity, orthogonal loss, reconstruction loss and

Model	Acc2(↑)	F1(↑)	MAE(↓)	Corr(↑)
TFN(B)[1]	80.8	80.7	0.901	0.698
LMF(B)[1]	82.5	82.4	0.917	0.695
MulT[1]	83.0	82.8	0.871	0.698
ICCN(B)[1]	83.0	83.0	0.860	0.710
MISA(B)*	82.8	82.7	0.784	0.748
MAG-BERT(B)*	83.6	83.6	0.782	0.763
MHMF-BERT(B)*	**85.3**	**85.3**	**0.748**	**0.787**

Table 1. The comparison with baselines on CMU-MOSI. (B) means the language features are based on BERT. [1] is from [7]. [2] is from [8]. Models with * are reproduced under the same conditions†.

Model	Acc2(↑)	F1(↑)	MAE(↓)	Corr(↑)
TFN(B)[1]	82.5	82.1	0.593	0.700
LMF(B)[1]	82.0	82.1	0.623	0.677
MulT[1]	82.5	82.3	0.580	0.703
ICCN(B)[1]	84.2	84.2	0.565	0.713
MISA(B)*	84.3	84.2	0.564	0.730
MAG-BERT(B)*	85.1	85.1	0.557	0.752
MHMF-BERT(B)*	**85.6**	**85.6**	**0.556**	**0.761**

Table 2. The comparison with baselines on CMU-MOSEI. (B) means the language features are based on BERT. [1] is from [7]. Models with * are reproduced under the same conditions†.

task prediction loss to learn modality-invariant and modality-specific representation.

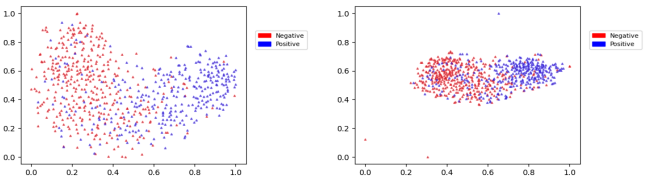
MAG-BERT (Multimodal Adaption Gate for BERT) [8] proposes an attachment called Multimodal Adaptation Gate that enables BERT and XLNet to accept multimodal data during fine-tuning.

4. RESULTS AND DISCUSSION

4.1. Results on comparable datasets

The comparative results on MOSI and MOSEI datasets are shown in Table 1,2. Compared with the baseline models using BERT word embedding, MHMF-BERT achieves the best performance. For the SOTA methods MISA and MAG-BERT, we reproduce them under the same conditions. It can be seen that our model surpasses them on all the metrics. Specifically, on CMU-MOSI dataset, our MHMF-BERT outperforms MAG-BERT by 1.7% on Acc2, surpasses MISA by 2.5% on Acc2; on CMU-MOSEI dataset, our proposed method is 0.5% higher than MAG and 1.3% higher than MISA on Acc2. These results prove that the effectiveness of our proposed model, indicating MHMF-BERT can better use BERT to model multimodal information, and our proposed new framework for fine-tuning Bert by using intermediate layer information is successful.

†The experimental environment is a WIN10 64-bit system, and the GPU is Nvidia Gefore RTX2080Ti. For a fair comparison, we run our model and SOTA models MISA, MAG-BERT on MOSI dataset five times, on MOSEI dataset three times, and report the average performances.



(a) Learned without restitution (b) Learned by MHMF-BERT

Fig. 4. T-SNE visualization of the learned features.

4.2. Fine-tuning Effect

In order to prove that our proposed method successfully fine-tunes BERT on MSA task, we set up three sets of experiments. Experiment 1 initializes weights of BERT randomly. This setting indicates that we do not use pretrained knowledge of BERT, only its network structure. Its performance on Acc2 for the CMU-MOSI is 71.9%. Experiment 2 freezes the parameters of all layers of BERT during training to avoid updating its pretrained weights. It means that we only use pretrained knowledge of BERT to extract text features and do not fine-tune it on multimodal data. This method can get an accuracy of 80.2% on Acc2. Experiment 3 does not impose any restrictions on the weights of pretrained BERT, and the weights will be fine-tuned during the training process. It gets an accuracy rate of 85.3% on Acc2. These results further confirm that MHMF-BERT uses the pretrained knowledge of the Bert middle layers and successfully fine-tunes BERT on multimodal data.

4.3. Visualization for the Learned Features

We provide visualization for distributions of learned multimodal features on Fig.4, which is obtained by using t-SNE. We observe that, (a) without restitution, the distribution of the features is very scattered, and the features of different classes do not form apparent clusters. (b) After adding the restitution module, constrained by the dual restitution loss, the distance between features of the same label becomes compact, and the distance between features of different labels becomes further apart, evident clustering appears, which is conducive to the prediction of the classifier. However, some data points that are difficult to distinguish are classified into the wrong clusters, which is reasonable because sentiment classification accuracy is about 85%.

4.4. Effect of Adaptation at Different Layers

We explore which layer of BERT fuse with nonverbal modal information will have better results. The results are shown in

Experiment	Audio	Vision	Acc2(↑)
1	{12}	{12}	83.9
2	{3,4}	{2,3}	84.0
3	{5,6}	{6,7}	84.4
4	{4,11}	{3,10}	85.1
5	{7,11}	{6,10}	85.3

Table 3. The results of applying CA modules to different layers of BERT. {2,3} means that the first CA module receives the output of the 2nd layer of BERT, and the second CA module receives the output of the 3rd layer of BERT.

Table 3. In the first set of experiments, we let CA receives output of last layer. The accuracy rate is not ideal. It proves the rationality and necessity of using BERT intermediate layer information for multimodal fusion. When CA is applied to the middle layers, we can see that when the two BERT layers connected to the CA modules are closer, the effect is often mediocre (experiment 2,3). We believe that this is because the information of BERT's adjacent layers is relatively similar, resulting in not providing more useful information. In addition, letting the CA modules connect to two layers that are far apart can positively impact the results. Because nonverbal modality features in the early modeling process can better match the phrase-level features of BERT's early layers. With the deepening of hierarchical fusion, more complex features with global dependencies will occur, fusing with high-level semantic information of BERT will be conducive.

5. CONCLUSION

This paper introduces a new way of using BERT to model multimodal information. We input the BERT intermediate layer information into a multimodal fusion structure external to BERT, where it integrates hierarchically with non-verbal modalities. In particular, our proposed model contains a restitution module, which can restore sentiment-relevant features from discarded information to the network, making the features more discriminative. Our model successfully fine-tuned BERT for multimodal language data. Its excellent performance on the two datasets of MOSI and MOSEI also proves that this is an effective multimodal fusion framework.

6. ACKNOWLEDGEMENT

This work was supported by two National Science Fund of China grants, No.61773215 and No.71974094.

7. REFERENCES

[1] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L. P. Morency, "Tensor fusion network for multimodal sentiment analysis," *arXiv*, 2017.

[2] Zhun Liu, Ying Shen, V. Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *ACL*, 2018.

[3] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," 2018.

[4] Yhh Tsai, S. Bai, P. P. Liang, J. Z. Kolter, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[5] D. Gkoumas, Q. Li, C. Lioma, Y. Yu, and D. Song, "What makes the difference? an empirical comparison of fusion strategies for multimodal language analysis," *Information Fusion*, vol. 66, no. 6, pp. 184–197, 2021.

[6] J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2019.

[7] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria, "Misa: Modality-invariant and -specific representations for multimodal sentiment analysis," *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

[8] Wasifur Rahman, M. Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and E. Hoque, "Integrating multimodal information in large pretrained transformers," *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2020, pp. 2359–2369, 2020.

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

[10] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[11] X. Jin, C. Lan, W. Zeng, and Z. Chen, "Style normalization and restitution for domaingeneralization and adaptation," 2021.

[12] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *ArXiv*, vol. abs/1706.03762, 2017.

[13] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intelligent Systems*, vol. 31, pp. 82–88, 2016.

[14] Amir Zadeh, Paul Pu Liang, Soujanya Poria, E. Cambria, and Louis-Philippe Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *ACL*, 2018.

[15] Zhongkai Sun, P. Sarma, W. Sethares, and Yingyu Liang, "Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis," in *AAAI*, 2020.