A Text-Centered Shared-Private Framework via Cross-Modal Prediction for Multimodal Sentiment Analysis

Yang Wu[1] **Zijie Lin**[1] **Yanyan Zhao**[1*] **Bing Qin**[1] **Li-Nan Zhu**[2]
[1] Harbin Institute of Technology [2] Zhejiang University of Technology
[1] {ywu, linzijie97, yyzhao, qinb}@ir.hit.edu.cn
[2] zln@zjut.edu.cn

Abstract

Multimodal fusion is a core problem for multimodal sentiment analysis. Previous works usually treat all three modal features equally and implicitly explore the interactions between different modalities. In this paper, we break this kind of methods in two ways. Firstly, we observe that textual modality plays the most important role in multimodal sentiment analysis, and this can be seen from the previous works. Secondly, we observe that comparing to the textual modality, the other two kinds of non-textual modalities (visual and acoustic) can provide two kinds of semantics, shared and private semantics. The shared semantics from the other two modalities can obviously enhance the textual semantics and make the sentiment analysis model more robust, and the private semantics can be complementary to the textual semantics and meanwhile provide different views to improve the performance of sentiment analysis together with the shared semantics. Motivated by these two observations, we propose a text-centered shared-private framework (TCSP) for multimodal fusion, which consists of the cross-modal prediction and sentiment regression parts. Experiments on the MOSEI and MOSI datasets demonstrate the effectiveness of our shared-private framework, which outperforms all baselines. Furthermore, our approach provides a new way to utilize the unlabeled data for multimodal sentiment analysis.

1 Introduction

Multimodal sentiment analysis is an emerging research field, which aims to understand people's sentiment using not only textual but also non-textual (visual, acoustic) data. This task has attracted increasing attention from the community recently, as people have realized that non-textual clues are helpful for detecting sentiment and the huge demands for the identification of opinions and sentiment in the video.

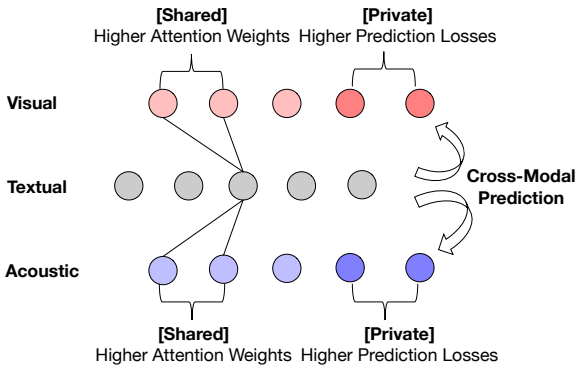

Figure 1: Distinguishing the shared and private features via cross-modal prediction.

Comparing to the traditional textual sentiment analysis(Liu, 2012), previous work demonstrates that the other non-textual data can improve the final performance (Chen et al., 2017; Zadeh et al., 2018b; Sun et al., 2020). There are two reasons. The first reason is that the three modalities can convey some common semantics. In this case, these non-textual common semantics do not provide additional information beyond textual data, but the repetitive information from them can strengthen the final performance. We call them *shared* semantics. The other reason is that the three modalities have their own special semantics, which are different to other modalities. These non-textual private semantics is modality-specific and hard to be predicted only by textual data. Thus this kind of semantics from the non-textual modalities can help to detect the final sentiment more accurately. We call them *private* semantics.

Previous works usually don't distinguish the shared semantics and the private semantics but treat each modal semantics as a whole, lacking the ability to explicitly explore the interaction between

* Corresponding Author

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4730–4738

different modalities. In this paper, we propose a text-centered shared-private framework for multimodal sentiment analysis. In this framework, the textual modality is considered as the core modality, and we first design a cross-modal prediction task to explicitly distinguish the shared and the private semantics between the textual modality and the non-textual(visual, acoustic) modality and then propose the sentiment regression model including the shared and private modules to fuse the textual features with two types of non-textual features.

In order to explore the shared and private semantics from non-textual modalities, we design the cross-modal prediction task, which is like a machine translation framework. The source is a sequence of textual modal features and the target is a sequence of another modal (visual or acoustic) features. We can explore the shared and private semantics via training two cross-modal prediction models, textual-to-visual and textual-to-acoustic models. In specific, as shown in Figure 1, we apply the pre-trained textual-to-visual and textual-to-acoustic models to predict the targets. The features of the target modality with higher prediction losses are distinguished as private. For each word, those putting higher attention weights on this word are distinguished as shared.

After obtaining shared and private features, we propose the sentiment regression model to fuse the textual features with two types of features. The sentiment regression model mainly consists of three parts, shared module, private module, and regression layer. In the shared module, each textual feature interacts with the corresponding shared features to get the enhanced features, which are then fed into a fusion block to obtain the final shared representation. Meanwhile, in the private module, the private features of each non-textual modality are passed through the attention layer to obtain the final private representation. Finally, we feed the obtained representations into the regression layer to predict the sentiment score.

We conduct experiments on two multimodal sentiment analysis benchmarks: CMU-MOSI and CMU-MOSEI. The experimental results show that our model outperforms all baselines. This can demonstrate that the shared-private framework for multimodal sentiment analysis can explicitly use the shared semantics between different modalities to enhance the final performance of sentiment analysis, and meanwhile can explicitly use the private semantics of each modality as the supplemental clues for sentiment analysis. In addition, we can observe that our designed cross-modal prediction task can accurately distinguish the shared and private non-textual semantics.

Our contributions can be concluded as follows.

1. We propose a challenging text-centered shared-private framework for the multimodal sentiment analysis. This framework can effectively fuse textual and non-textual features benefitting from the unlabeled data.

2. We design a cross-modal prediction task to explore the shared and private semantics for each non-textual modality.

3. We achieve significant improvements from learning the shared and private semantics from different modalities compared to the algorithms without distinguishing the shared and private semantics.

2 Related Work

There are two lines of works conducted on multimodal sentiment analysis.

One is focusing on utterance-level multimodal feature fusion. These methods use the features of the overall utterance. For example, they first extract the frame-level visual or acoustic features and then average them to obtain the final features, which are called utterance-level features. The utterance-level textual features can be obtained by applying RNNs for words. The obtained utterance-level features are fed into the fusion model to get the multimodal representation. Some models have been proposed for effective multimodal feature fusion. Zadeh et al. (2017) proposed Tensor Fusion to explicitly capture unimodal, bimodal, and trimodal interactions. But this method uses the three-fold Cartesian product to fuse the multimodal features, which makes the time cost very high. To address it, Liu et al. (2018) presented the Efficient Low-rank Multimodal Fusion, which applies multimodal fusion using low-rank tensors to accelerate the fusion process. Mai et al. (2020) proposed a graph fusion network to model unimodal, bimodal, and trimodal interactions successively.

The utterance-level features mainly contain global information, which may fail to capture local information. Therefore, recent works are mostly focusing on word-level multimodal feature fusion.

And our work in this paper is also based on word-level features. To extract word-level features, the first step is applying force alignment to obtain the timestamps of each word including the start time and end time. And then following the timestamps, the utterance is split into some video clips. Finally, word-level visual or acoustic features are obtained by averaging the frame-level features of the video clips. Based on word-level features, lots of methods are proposed for performing word-level multimodal feature fusion. Zadeh et al. (2018a) proposed the Memory Fusion Network(MFN) to capture the interactions across both different modalities and timesteps. Inspired by the observation that the meaning of words often varies dynamically in different non-verbal contexts, Wang et al. (2019) proposed the Recurrent Attended Variation Embedding Network (RAVEN). This model applies the Attention Gating module to fuse the word-level features, which can dynamically use the non-verbal features to shift the word embeddings. Tsai et al. (2019) presented multimodal transformer (Mult), which uses the cross-modal attention to capture the bimodal interactions, motivated by the great success of transformer in NLP(Vaswani et al., 2017).

Besides, there is a related work (Pham et al., 2019) need to be noticed, which proposed that translation from a source to a target modality provides a way to learn joint representations and proposed the Multimodal Cyclic Translation Network model (MCTN) to learn joint multimodal representations. Comparing to this work, our model has a significant difference. That is we use the cross-modal prediction task to distinguish the shared and private non-textual features instead of training the model as an auxiliary task. In this way, we can obtain more useful information by deeply probing the cross-modal prediction model.

3 Approach

In this section, we will introduce our shared-private framework for multimodal sentiment analysis in detail. In this framework, we treat the textual modality as the core, then how to explore the shared and the private semantics of the non-textual modality compared to the textual modality, and how to fuse all three modal features are two important steps. For the first step, we design a cross-modal prediction task and obtain the shared and private features of the non-textual modalities via training two cross-modal prediction models, textual-to-visual and textual-to-acoustic models. And for the second step, we design a sentiment regression model to fuse the textual features and the two types of features.

3.1 Cross-Modal Prediction

Task Definition: The cross-modal prediction task is formalized as follows. Given a sequence of textual features denoted as $\mathbf{x}_l = \{x_l^t : 1 \leq t \leq L, x_l^t \in \mathbb{R}^{d_l}\}$, L is the length of the given sequence, t is the timestep, and the goal is to predict the corresponding sequence of visual or acoustic features denoted as $\mathbf{x}_i = \{x_i^t : 1 \leq t \leq L, x_i^t \in \mathbb{R}^{d_i}\}$, $i \in \{v, a\}$. Cross-modal prediction task is inspired by the machine translation task. The inputs are the textual features, and the generated outputs are the non-textual (visual or acoustic) features. During the translation from the textual modality to other modalities, we can exploit the shared and the private semantics of the non-textual modalities.

Prediction Model: We use the Seq2Seq model with attention (Bahdanau et al., 2015) as our model framework. The encoder takes the textual features $\mathbf{x}_l$ as inputs and outputs the hidden states $\mathbf{h}_{enc} = \{h_{enc}^t : 1 \leq t \leq L, h_{enc}^t \in \mathbb{R}^{d_h}\}$. The decoder takes the previous hidden state h_{dec}^{t-1} and hidden states of the encoder as inputs and predicts the non-textual feature x_i^t, $i \in \{v, a\}$, at the t timestep. We choose the MSE as our loss function. The prediction loss values are denoted as $\mathbf{e}_{l \to i} = \{e_{l \to i}^t : 1 \leq t \leq L\}$. The attention map of the prediction model is denoted as $\mathbf{m}_{i \to l}$. In practice, we apply LSTMs (Hochreiter and Schmidhuber, 1997) to implement our encoders and decoders. After training the cross-modal prediction models using textual-visual and textual-acoustic paired data, we can obtain two models, textual-to-visual and textual-to-acoustic models. We then use the obtained models to distinguish the shared and private features and record the results as shared and private masks, which will be passed to the sentiment regression model.

Shared Mask: We propose the shared mask to find out the shared semantics of the two kinds of non-textual modalities. The basic assumption is that during cross-modal prediction, if the prediction model wants to generate a non-textual feature as precisely as possible, it should pay more attention to the input textual features, that contain more shared semantics. Based on this assumption, we design the method to obtain the shared mask. Given

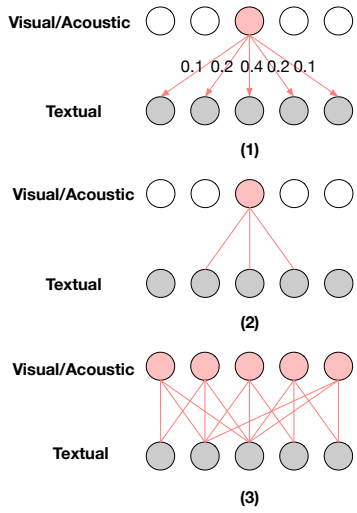

Figure 2: Obtaining the shared mask from the cross-modal prediction model.

$\mathbf{m}_{i\to l}$, $i \in \{v, a\}$, for each row t, we first sort the attention weights $\mathbf{m}_{i\to l}^{t,*}$ and then get the indexes S^t of the largest K_s values. Finally, we can get the shared mask $smask$, $smask \in \mathbb{R}^{L*L}$. $smask^{t_1,t_2}$ is 1 if the $t_1 \in S^{t_2}$ and 0 otherwise.

To describe this method intuitively, we show this process in Figure 2. There are three steps: (1) We build an attention graph and the values of edges mean the attention weights. We illustrate a part of it for simplicity. (2) We only keep the edges with larger values for each non-textual node and delete others. (3) We map the graph to the shared mask $smask$, $smask^{t_1,t_2}$ is 1 if there is a edge between textual node t_1 and non-textual node t_2 and 0 otherwise. The shared mask will be passed to the share module of the sentiment regression model to make the model focus on the shared features.

Private Mask: In order to find out the private semantics of the two kinds of non-textual modalities, we propose the private mask. The basic assumption is that the features containing modality-private information are difficult to be predicted by textual modality. The private mask of a given utterance is obtained as follows. Given an utterance which includes three modalities, textual, visual and acoustic, denoted as $\mathbf{x}_i = \{x_i^t : 1 \le t \le L, x_i^t \in \mathbb{R}^{d_i}\}$, $i \in \{l, v, a\}$, we first use the trained prediction models to get the loss values, $\mathbf{e}_{l\to v}$ and $\mathbf{e}_{l\to a}$. Then we sort the loss values to obtain the indexes P of the largest K_p values. Finally, We can get the private mask $pmask$, $pmask \in \mathbb{R}^L$. $pmask^t$ is 1 if the $t \in P$ and 0 otherwise. The private mask will be used by the private module of the sentiment regression model to force the model to focus on private features.

3.2 Regression

In this section, we study how to fuse the shared and private information obtained from Section 3.1. An illustration of our framework is given in **Figure 3**.

3.2.1 Input Layer

Given an utterance which includes three modalities, textual, visual and acoustic, the extracted multimodal features are denoted as $\mathbf{x}_i = \{x_i^t : 1 \le t \le L, x_i^t \in \mathbb{R}^{d_i}\}$, $i \in \{l, v, a\}$. We use three LSTM networks to encode the input multimodal features $\mathbf{x}_i$, producing $\mathbf{h}_i = \{h_i^t : 1 \le t \le L, h_i^t \in \mathbb{R}^{d_h}\}$.

$$\begin{aligned} \mathbf{h}_l &= \mathrm{LSTM_l}(\mathbf{x_l}) \\ \mathbf{h}_v &= \mathrm{LSTM_v}(\mathbf{x_v}) \\ \mathbf{h}_a &= \mathrm{LSTM_a}(\mathbf{x_a}) \end{aligned} \tag{1}$$

3.2.2 Shared Module

The core idea of the shared module is leveraging the shared information from non-textual modal features to enhance the representations of words. To achieve it, we propose the masked cross-modal attention network, which can utilize the shared masks obtained from cross-modal prediction models to focus on the non-textual shared features.

In the masked cross-modal attention network, we first calculate the attention scores across the non-textual representations $\mathbf{h}_i$, $i \in \{v, a\}$, for each word. We denote the scores as $\mathbf{s}_{l\to i}$.

$$\begin{aligned} \mathbf{s}_{l\to v}^{t_1,t_2} &= W_2(tanh(W_1([h_l^{t_1}; h_v^{t_2}]) + b_1)) \\ \mathbf{s}_{l\to a}^{t_1,t_2} &= W_4(tanh(W_3([h_l^{t_1}; h_a^{t_2}]) + b_3)) \end{aligned} \tag{2}$$

where $W_1, W_3 \in \mathbb{R}^{d_h \times 2d_h}$, $W_2, W_4 \in \mathbb{R}^{1\times d_h}$, $b_1, b_3 \in \mathbb{R}^{d_h}$ are the parameters of the score functions.

To focus on the shared features, we first calculate the attention weights $\mathbf{w}_{l\to v}$ and $\mathbf{w}_{l\to a}$ using the softmax function and mask the other features out using the shared mask.

$$\begin{aligned} \mathbf{w}_{l\to v}^{t_1,t_2} &= \frac{e^{\mathbf{s}_{l\to v}^{t_1,t_2}}}{\sum_{t_3=1}^{L} e^{\mathbf{s}_{l\to v}^{t_1,t_3}}} \\ \mathbf{w}_{l\to a}^{t_1,t_2} &= \frac{e^{\mathbf{s}_{l\to a}^{t_1,t_2}}}{\sum_{t_3=1}^{L} e^{\mathbf{s}_{l\to a}^{t_1,t_3}}} \end{aligned} \tag{3}$$

$$\begin{aligned} \mathbf{w}_{l\to v} &= \mathbf{w}_{l\to v} \circ smask_{l\to v} \\ \mathbf{w}_{l\to a} &= \mathbf{w}_{l\to a} \circ smask_{l\to a} \end{aligned} \tag{4}$$

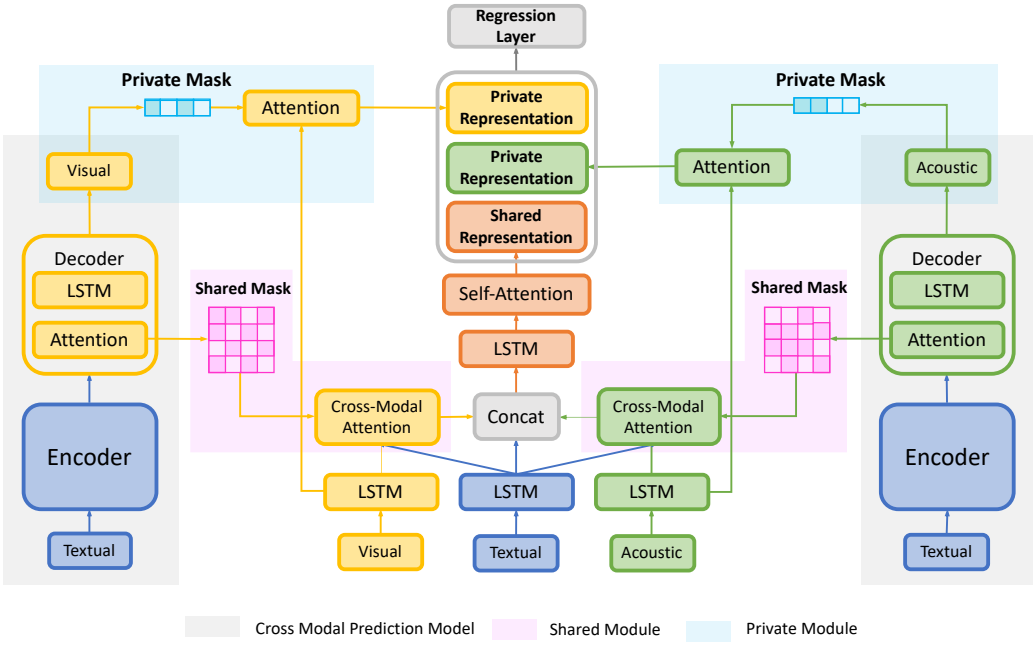

Figure 3: Illustration of our shared-private framework.

We obtain the non-textual shared context vectors $\mathbf{c}_v$ and $\mathbf{c}_a$. $\mathbf{c}_v, \mathbf{c}_a \in \mathbb{R}^{L \times d_h}$.

$$\begin{aligned} \mathbf{c}_v &= \mathbf{w}_{l \to v} \mathbf{h}_v \\ \mathbf{c}_a &= \mathbf{w}_{l \to a} \mathbf{h}_a \end{aligned} \tag{5}$$

To fuse textual and non-textual shared features, we concatenate $\mathbf{c}_v$, $\mathbf{c}_a$, and $\mathbf{h}_l$ and feed it into the fusion LSTM network, producing $\mathbf{r}_s \in \mathbb{R}^{L \times 3d_h}$. We further use a self-attention layer, which is denoted as $\mathrm{SelfAttentionLayer}$, to learn the final representation. The self-attention layer is similar to the cross-modal attention network. We use the last step representation of $\mathbf{r}_n$ as the shared representation, which is denoted as r_s.

$$\begin{aligned} \mathbf{r}_m &= \mathrm{LSTM}_{\mathrm{fusion}}([\mathbf{c}_\mathrm{v}; \mathbf{c}_\mathrm{a}; \mathbf{h}_\mathrm{l}]) \\ \mathbf{r}_n &= \mathrm{SelfAttentionLayer}(\mathbf{r}_\mathrm{m}) \end{aligned} \tag{6}$$

3.2.3 Private Module

To enable the model to capture the unique information contained in non-textual modalities, we design the private module. Specifically, we use the attention network to learn informative and modality-private representations.

$$\begin{aligned} \mathbf{s}_v^t &= W_5 h_v^t + b_5 \\ \mathbf{s}_a^t &= W_6 h_a^t + b_6 \end{aligned} \tag{7}$$

where $W_5, W_6 \in \mathbb{R}^{1 \times d_h}$, $b_5, b_6 \in \mathbb{R}$ are the parameters of the score functions.

We use private masks to ignore other features and apply the softmax function to get the attention weights.

$$\begin{aligned} \mathbf{s}_v &= \mathbf{s}_v + (1 - pmask_v) * (-10^8) \\ \mathbf{s}_a &= \mathbf{s}_a + (1 - pmask_a) * (-10^8) \end{aligned} \tag{8}$$

Finally, we compute the weighted sum and represent them as p_v and p_a, which are called the private representations.

$$\begin{aligned} \mathbf{w}_v^t &= \frac{e^{\mathbf{s}_v^t}}{\sum_{t_1=1}^{L} e^{\mathbf{s}_v^{t_1}}} \\ \mathbf{w}_a^t &= \frac{e^{\mathbf{s}_a^t}}{\sum_{t_1=1}^{L} e^{\mathbf{s}_a^{t_1}}} \end{aligned} \tag{9}$$

$$\begin{aligned} p_v &= \mathbf{w}_v \mathbf{h}_v \\ p_a &= \mathbf{w}_a \mathbf{h}_a \end{aligned} \tag{10}$$

3.2.4 Regression Layer

We design the regression layer, which is implemented by a two-layer network with ReLU activation function, to fuse the shared and private representations.

$$\hat{y} = W_o(ReLU(W_f([r_s; p_v; p_a]) + b_f)) + b_o \tag{11}$$

where $W_f \in \mathbb{R}^{d_h \times 5d_h}$, $W_o \in \mathbb{R}^{1 \times d_h}$, $b_f \in \mathbb{R}^{d_h}$, $b_o \in \mathbb{R}$.

Table 1: Hyperparameters of our model.

Models	Parameters	MOSI	MOSEI
Cross-Modal Prediction	Batch Size	24	24
	Max Length	50	128
	Hidden Size	100	100
	Epochs	40	40
	Learning Rate	0.0001	0.0001
	Dropout	0.5	0.5
	Patience	5	10
Regression	Batch Size	24	24
	Max Length	50	128
	Hidden Size	100	100
	Epochs	30	30
	Learning Rate	0.001	0.001
	Dropout	0.5	0.5
	Selection Number	5	5
	Patience	5	5

4 Experiments

4.1 Datasets

We conduct experiments on two public datasets, CMU-MOSI (Zadeh, 2015) and CMU-MOSEI (Zadeh et al., 2018b) to evaluate our proposed model. CMU multimodal opinion-level sentiment intensity (CMU-MOSI) consists of 93 videos collected from the YouTube website. The length of the videos varies from 2-5 mins. These videos are split into 2,199 short video clips and labeled with sentiment scores from -3 (strongly negative) to 3 (strongly positive). CMU multimodal opinion sentiment and emotion intensity (CMU-MOSEI) consists of 23,453 annotated video utterances from 1,000 distinct speakers and 250 topics. Each utterance is annotated with sentiment scores from -3 (strongly negative) to 3 (strongly positive).

The multimodal features used in our experiments are described as follows. We use glove word embeddings (Pennington et al., 2014) to represent the words. The dimension of each word embedding is 300. We extract the visual features using Facet, which can extract 35 facial action units (Ekman et al., 1980; Ekman, 1992) from each frame resulting in a 35-dimensional vector. The acoustic features are obtained by applying COVAREP (Degottex et al., 2014), which includes 12 Mel-frequency cepstral coefficients (MFCCs) and other low-level features. The dimension of the acoustic feature is 74.

4.2 Evaluation Metrics

Following previous works, we take 2-class accuracy(Acc), f1 score(F1), mean absolute error (MAE), and correlation(Corr) as our evaluation metrics. As the prediction results are real values, we first use mean absolute error and Corr between prediction scores and ground truths to evaluate the models. In addition, we then map the sentiment scores into sentiment labels and use classification metrics, such as accuracy and f1 score, to assess the model performance.

4.3 Training Details

The hyperparameters of our model are listed in Table 1. In practice, we apply dropout before the last linear layer for regularization. We use Adam as the optimizer. The learning rate is decayed once the validation loss stops decreasing. The Selection Number is the number of selected shared/private features, K_s and K_p. We take the same value for K_s and K_p for simplicity.

4.4 Baselines

We compare our proposed model with the following baselines. **EF-LSTM** fuses the multimodal features by concatenating and applies an LSTM network to get the final representation. **LF-LSTM** first uses three LSTM networks to encode three modal features and concatenates three obtained representations to get the final representation. **MFN** (Zadeh et al., 2018a) captures the interactions across both the different modalities and time. **RAVEN** (Wang et al., 2019) first combines the nonverbal information with word representations and then feeds the modified word representations into an LSTM network to obtain the utterance representation. **MCTN** (Pham et al., 2019) learns joint multimodal representations by translating between modalities. **MulT** (Tsai et al., 2019) uses crossmodal transformers to fuse multimodal features. **Multimodal Routing** (Tsai et al., 2020) proposes a routing mechanism to capture the interactions between input modalities and outputs. **TCSP(Base)** is our base model. The model architecture is as same as our full model, but it doesn't use shared and private masks. Comparing TCSP(Base) and TCSP(Full), we can judge whether distinguishing the shared and private features of non-textual modalities is useful.

4.5 Experimental Results

We compare our model with several baselines and the experimental results are shown in Table 2. Comparing our base model with other baselines, our base model fails to obtain the best result and underperforms RAVEN and MulT for the Acc, F1

Table 2: Experimental results on the test sets of the MOSEI and MOSI dataset. The best results are in **bold**. As the Multimodal Routing model is designed for classification, we don't report the regression metrics of it for fair comparison.

Models	MOSI				MOSEI			
	Acc ↑	F1 ↑	MAE ↓	Corr ↑	Acc ↑	F1 ↑	MAE ↓	Corr ↑
EF-LSTM	76.0/75.3	75.9/75.2	1.020/1.023	0.603/0.608	78.4/78.2	79.5/77.9	0.642/0.642	0.641/0.616
LF-LSTM	75.3/76.8	75.1/76.7	1.046/1.015	0.600/0.625	80.3/80.6	80.8/80.6	0.606/0.619	0.676/0.659
MFN	74.5/77.4	74.4/77.3	1.036/0.965	0.607/0.632	78.1/-	79.2/-	0.640/-	0.637/-
RAVEN	76.2/78.0	76.0/76.6	1.012/0.915	0.614/0.691	81.3/79.1	81.6/79.5	0.595/0.614	0.701/0.662
MCTN	71.6/79.3	71.5/79.1	1.142/0.909	0.487/0.676	80.8/79.8	80.6/80.6	0.611/0.609	0.670/0.670
MulT	78.9/83.0	78.8/82.8	1.000/0.871	0.670/0.698	81.8/82.5	81.8/82.3	0.605/0.580	0.682/0.703
Multimodal Routing	68.5/-	68.4/-	-/-	-/-	76.0/81.7	75.6/81.8	-/-	-/-
TCSP(Base)	79.3	79.3	0.956	0.658	80.7	80.3	0.593	0.692
TCSP(Full)	**80.9**	**81.0**	**0.908**	**0.710**	**82.8**	**82.7**	**0.576**	**0.715**

Table 3: Ablation analysis of TCSP evaluated on the test data. The best results are in **bold**.

Models	MOSI				MOSEI			
	Acc ↑	F1 ↑	MAE ↓	Corr ↑	Acc ↑	F1 ↑	MAE ↓	Corr ↑
TCSP	**80.9**	**81.0**	**0.908**	**0.710**	**82.8**	**82.7**	**0.576**	**0.715**
w/o Private Mask	79.9	79.8	0.930	0.663	82.2	82.1	0.576	0.710
w/o Shared Mask	79.0	79.0	0.965	0.660	82.3	82.1	0.585	0.701
w/o Both Masks	79.3	79.3	0.956	0.658	80.7	80.3	0.593	0.692

metrics on the MOSEI dataset. However, with the help of the cross-modal prediction task, our text-centered shared-private framework (TCSP) achieves the best performance and outperforms all baselines on both datasets. This can demonstrate that the shared-private framework proposed in this paper is effective for multimodal sentiment analysis. Furthermore, it can be observed that the shared and private features for each non-textual modality obtained from the cross-modal prediction task can provide useful clues for the interactions between different modalities. Thus, these non-textual shared-private features can be jointly fused with the textual features to improve the multimodal sentiment analysis.

We also observe that there is a larger margin between our full model and our base model on the MOSI dataset. We attribute it to the small data size of the MOSI dataset. It is insufficient for training the base model, which makes it benefit more from the shared and private information.

It should be noted that, in Table 2, we provide two results for each method on each dataset. The left result is obtained by rerunning the public codes in the same experimental setting, which refers to the same dataset split and the same extracted features of three modalities. The right result is copied from previous papers and the experimental settings are different. To guarantee the justice, we compare our TCSP model with the left results.

5 Analysis

5.1 Ablation Study

We conduct the ablation experiments to distinguish the contribution of each part. As shown in Table 3, ablating either shared mask or private mask hurts the model performance, which indicates that both masks are useful for the sentiment prediction. The shared mask can enable the sentiment regression model to get the modality-shared features resulting in a more robust regression model. The private mask makes the regression model focus on modality-private features, which provides extra information for sentiment prediction. With the help of shared and private masks, the regression model in the shared-private framework can fuse the textual features with two types of non-textual features individually, which is the more effective method for multimodal feature fusion.

5.2 Effect of Selection Number

Selection Number is the number of selected shared/private features, K_s and K_p. We take the same value for K_s and K_p for simplicity. We evaluate our model with different selected numbers from 1 to 8 on the MOSEI dataset to quantify the effect. The experimental results are shown in Figure 4. We can observe that our model achieves the best performance on the Acc and F1 metrics when the Selection Number is 5. The possible reason is that

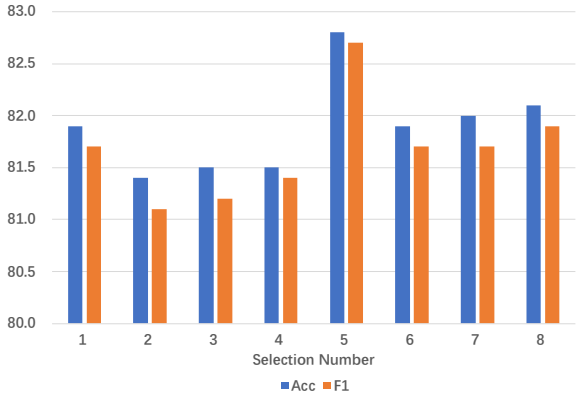

Figure 4: Experimental results with different selection numbers on the MOSEI dataset.

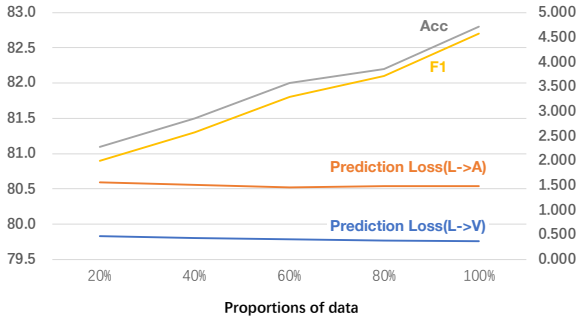

Figure 5: Experimental results on the MOSEI dataset with different proportions of data used for the cross-modal prediction model.

too small Selection Number makes the model only focus on few features. This could result in missing useful information. In contrast, too large one makes the model attend too many features, which weakens the effect of masks. For this reason, selecting a middle number could be better.

5.3 Effect of Cross-Modal Prediction Model

The cross-modal prediction task is the core of our shared-private framework, and it has been demonstrated that this task is effective from Table 3 and Section 5.1. In this section, we want to further explore the effect of cross model prediction for the final regression model.

In Figure 5, we design different cross-modal prediction models trained with different proportions (from 20% to 100%) of MOSEI data and then fuse the obtained shared and private information into the regression models. It should be noticed that all regression models are trained with all data of the MOSEI dataset. The results show that when we use more data, the final performance is better. And meanwhile, it can be observed that the two kinds of prediction losses (from textual to visual modality and from textual to acoustic modality) are decreased when the proportion of the used data is increased.

This can reveal that the cross-modal prediction model trained with more data can provide more informative supervision signals, which are the shared and private masks specifically. If the performance of cross-modal prediction model is low, it is impossible to teach the regression model to play the precise role in the shared-private framework.

6 Conclusion

In this paper, we propose a text-centered shared-private framework for multimodal sentiment analysis. In this framework, we treat the textual modality as the core and aim to use the other non-textual modalities to help enrich the semantics of the textual modality. For each non-textual modality, we consider two types of semantics, shared and private, which have different functions. Shared semantics can enhance the textual semantics to make the model more robust and the private semantics can provide extra information for more precise prediction.

To distinguish these two semantics, we design a cross-modal prediction task and record the results as share and private masks. We further propose a regression model utilizing the shared and private modules to fuse the textual features with two non-textual features. The experimental results demonstrate that distinguishing the shared and private non-textual semantics and explicitly modeling the interactions between textual and two non-textual semantics is a better way for the multimodal sentiment analysis than just treating each non-textual features as a whole. The analyses show that the regression model can benefit more from the better cross-modal prediction model, which also indicates that the cross-modal prediction process can produce useful supervision signals only using unlabeled data.

In future work, we plan to collect more unlabeled data to enhance our model. In addition, we would also like to explore other approaches using the unlabeled data to help multimodal feature fusion.

Acknowledgments

This work was supported in part by the following Grants: National Natural Science Foundation of China (No. 61632011, No. 61772153), National Key R&D Program of China (No. 2018YFB1005103).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 163–171.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. Covarep - a collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 960–964. IEEE.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Paul Ekman, Wallace V Freisen, and Sonia Ancoli. 1980. Facial signs of emotional experience. *Journal of personality and social psychology*, 39(6):1125.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2247–2256.

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.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6892–6899.

Zhongkai Sun, Prathusha Sarma, William Sethares, and Yingyu Liang. 2020. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8992–8999.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569.

Yao-Hung Hubert Tsai, Martin Ma, Muqiao Yang, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Multimodal routing: Improving local and global interpretability of multimodal language analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1823–1833.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7216–7223.

Amir Zadeh. 2015. Micro-opinion sentiment intensity analysis and summarization in online videos. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 587–591.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark. Association for Computational Linguistics.

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

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018b. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246.