Token-Level Contrastive Learning with Modality-Aware Prompting for Multimodal Intent Recognition

Qianrui Zhou[1,2], **Hua Xu**[1,2*], **Hao Li**[1,2], **Hanlei Zhang**[1,2], **Xiaohan Zhang**[1,3], **Yifan Wang**[1,3], **Kai Gao**[3]

[1]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[2]Beijing National Research Center for Information Science and Technology (BNRist), Beijing 100084, China
[3]School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang 050018, China
zgr22@mails.tsinghua.edu.cn, xuhua@tsinghua.edu.cn

Abstract

Multimodal intent recognition aims to leverage diverse modalities such as expressions, body movements and tone of speech to comprehend user's intent, constituting a critical task for understanding human language and behavior in real-world multimodal scenarios. Nevertheless, the majority of existing methods ignore potential correlations among different modalities and own limitations in effectively learning semantic features from nonverbal modalities. In this paper, we introduce a token-level contrastive learning method with modality-aware prompting (TCL-MAP) to address the above challenges. To establish an optimal multimodal semantic environment for text modality, we develop a modality-aware prompting module (MAP), which effectively aligns and fuses features from text, video and audio modalities with similarity-based modality alignment and cross-modality attention mechanism. Based on the modality-aware prompt and ground truth labels, the proposed token-level contrastive learning framework (TCL) constructs augmented samples and employs NT-Xent loss on the label token. Specifically, TCL capitalizes on the optimal textual semantic insights derived from intent labels to guide the learning processes of other modalities in return. Extensive experiments show that our method achieves remarkable improvements compared to state-of-the-art methods. Additionally, ablation analyses demonstrate the superiority of the modality-aware prompt over the handcrafted prompt, which holds substantial significance for multimodal prompt learning. The codes are released at https://github.com/thuiar/TCL-MAP.

Introduction

Intent recognition is a pivotal component in natural language understanding(NLU), which is employed to classify intent categories in goal-oriented scenarios based on text information. Prior studies have extensively researched intent recognition and validated the significance of textual modality (Zhang et al. 2021a,b). Recently, multimodal intent recognition has been devised to analyze human intents by incorporating both natural language and other nonverbal information (e.g. video and audio). Contrary to depending on single modality, leveraging multiple modalities can provide a substantial amount of information, providing a distinct advantage in accurately identifying more intricate intent categories. Pioneering works in this field (Zhang et al. 2022; Saha et al. 2020) gather multimodal data from the real world to create intent recognition datasets, which affirms the crucial role of multimodal information on intent recognition tasks.

To effectively leaverage the data from various modalities, numerous methods have been proposed for multimodal language understanding. As the state-of-the-art methods for multimodal intent recognition, (Tsai et al. 2019; Hazarika, Zimmermann, and Poria 2020; Rahman et al. 2020) utilize Transformer-based techniques to integrate information from different modalities into a unified feature. Another series of works primarily focus on achieving significant representations by addressing the correlations among modalities. For instance, (Dong et al. 2022) introduce a framework for cross-modality contrastive learning aimed at narrowing the gap between modalities and enhancing multimodal representations. Moreover, the latest work (Yu et al. 2023) aims to incorporate knowledge from large-scale pre-trained models and has demonstrated remarkable progress.

However, existing approaches are adept at capturing basic semantics such as emotions and encounter limitations in capturing the deep semantic information inherent in intents using multimodal data, owing to deficiencies in managing potential correlations across distinct modalities and extracting high-quality semantic features from nonverbal modalities.To achieve a deeper understanding of human intents in the real world, there are two main challenges. Firstly, given that intent recognition is primarily a text-centric task, it is crucial to explore the association between nonverbal modalities and text modality. Secondly, mining semantic information from video and audio modality poses a significant difficulty due to the complex nature of the intent concept.

In this paper, we propose a Token-Level Contrastive Learning with Modality-Aware Prompting (TCL-MAP) approach, which generates prompts based on video and audio modalities to enhance the text representation and in return utilizes the high-quality textual feature to guide other modalities in learning semantic features. To tackle the first challenge, we design a modality-aware prompting (MAP) module to align different modalities based on their similarity and generate modality-aware prompt using a cross-modality attention mechanism. The generated prompt combines video and audio information with learnable parameters to establish a foundational semantic environment for text representation learning,

*Hua Xu is the corresponding author.

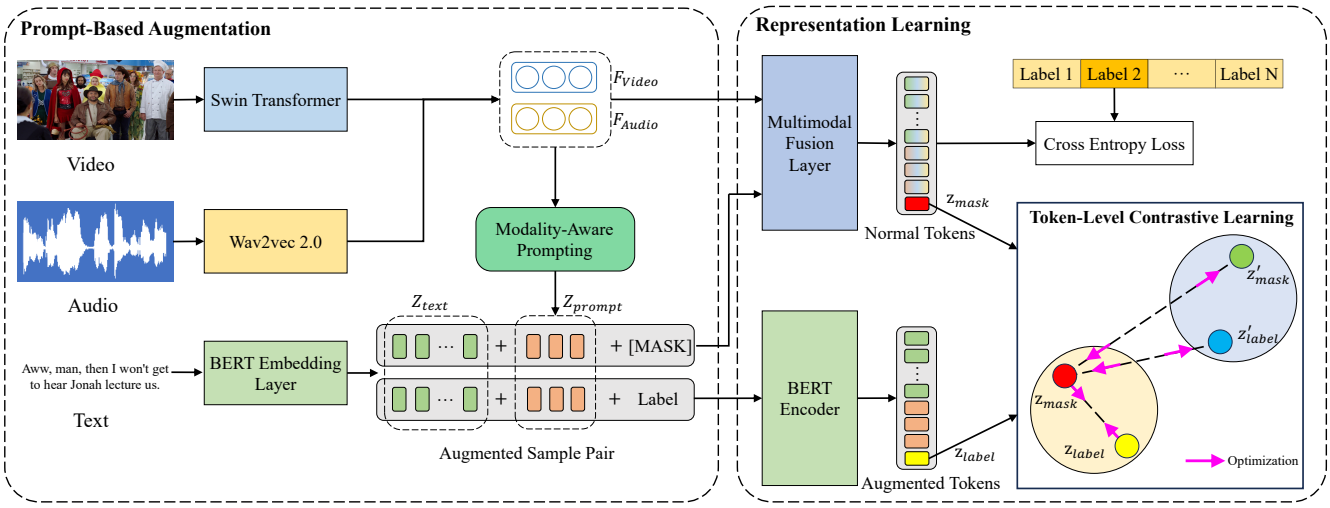

Figure 1: The overview architecture of TCL-MAP. In the Prompt-Based Augmentation module, we first create the modality-aware prompt using multimodal features, and then concatenate text tokens, prompt tokens and [MASK]/Label token to construct augmented pair. In the Representation Learning module, we extract the refined tokens for classification and conduct contrastive learning between the [MASK] token and the Label token.

mitigating the need for labor-intensive prompt engineering. For the second challenge, we utilize a token-level contrastive learning framework (TCL) which leverages premium supervised signals to facilitate the extraction of semantic information. Specifically, TCL initially constructs augmented samples by concatenating the modality-aware prompt and ground truth labels with the original text. Subsequently, TCL extracts the ground truth token from BERT (Devlin et al. 2018) to guide the process of semantic feature learning through the application of NT-Xent (Sohn 2016) loss.

Extensive experiments are performed on MIntRec (Zhang et al. 2022) and MELD-DA (Saha et al. 2020) datasets, which demonstrates the significant improvements achieved by our methods over the state-of-the-art methods. Moreover, ablation analyses reveal that the modality-aware prompt generated by MAP outperforms the handcrafted prompt, thereby contributing to the advancement of multimodal prompt learning. Our contributions can be summarized as follows:

- We design a modality-aware prompting module to generate modality-aware prompts using video and audio modalities, which establishes an optimal multimodal semantic context for text modality.
- We propose a token-level contrastive learning framework for leveraging semantic information from the ground truth label to guide other modalities in learning semantic representations. To the best of our knowledge, this is the first attempt that utilizes prompt learning to create premium supervised signals for contrastive learning.
- Comprehensive experiments conducted on two challenging datasets show that the proposed method achieves state-of-the-art performance on the multimodal intent recognition task.

Related Works

Multimodal Fusion Methods

Multimodal fusion techniques strive to achieve high-quality multimodal representations through effective fusion processes. Traditional approaches (Zadeh et al. 2017; Liu et al. 2018; Hou et al. 2019) harnesses the representational capabilities of tensors for multimodal data representation, showcasing their representational capabilities. However, these methods face the challenge of striking a balance between computational complexity and the quality of tensor representation. To solve the problem, MFN (Zadeh et al. 2018) first learns view-specific interactions and leverages an attention mechanism with a multi-perspective gated memory.

Recent methods are based on Transformer for multimodal fusion. By incorporating attention mechanism, MulT (Tsai et al. 2019) addresses the challenge of non-aligned multi-modal sequences and MISA (Hazarika, Zimmermann, and Poria 2020) aims to learn representations that exhibit both modality-invariant and modality-specific characteristics. Moreover, MAG-BERT (Rahman et al. 2020) introduces an attachment mechanism to enhance the fine-tuning process of BERT (Devlin et al. 2018). Subsequently, researchers start to prioritize the interactions within individual modalities. For example, MMIM (Han, Chen, and Poria 2021) focuses on the correlations between unimodal inputs and multimodal fusion features through maximizing the mutual information. To separately address the relationships between each pair of modalities, BBFN (Han et al. 2021) integrates two bimodal fusion modules along with a gated control mechanism. Given that top-down interactions remain unaccounted in previous approaches, MMLatch (Paraskevopoulos, Georgiou, and Potamianos 2022) addresses this limitation by incorporating a feedback mechanism in the forward pass.

Prompt Learning

Prompt learning originates from nature language processing (NLP) and is adopted to elicit useful information from pre-trained language models for downstream tasks. CoOp (Zhou et al. 2022b) first employ prompt learning in adapting large vision-language models, garnering the interest of numerous researchers. Subsequently, CoCoOp (Zhou et al. 2022a) addresses the weak generalizability issue of CoOp by generating an input-conditional token for each sample. To further incorporate contextual information into the prompt, DenseCLIP (Rao et al. 2022) introduces a transformer encoder within the CoCoOp framework to enhance the correlation between image embeddings and prompt tokens. Recent works (Wang et al. 2022; Li et al. 2022; Gan et al. 2023) have applied prompt learning to various domains in computer vision. Different from prior approaches, our method pioneers the application of prompt learning on multimodal tasks without reliance on extensive pre-trained models. Moreover, our research demonstrates the substantial impact of prompt learning in enhancing multimodal representations.

Contrastive Learning

Contrastive Learning emphasizes similarities and differences between data pairs, pulling similar pairs closer and pushing dissimilar pairs apart. Early approaches (Wu et al. 2018; Ye et al. 2019; Tian, Krishnan, and Isola 2020) lays the diverse foundation for contrastive learning's evolution, encompassing various models, loss functions, and pretext tasks. Subsequently, MoCo (He et al. 2020) views contrastive learning as dictionary look-up and exploits a dynamic dictionary with a queue and a momentum encoder. Notably, SimCLR (Chen et al. 2020) proposes a simple framework for contrastive learning using diverse data augmentation operations and a nonlinear projection head.

Recent methodologies aim to eliminate the necessity for negative examples and improve the efficacy of feature extraction. To enhance robustness in the presence of diverse augmentations, BYOL (Grill et al. 2020) proposes a slow-moving average target network output using the online network's output, effectively avoiding the need for negative pairs. Taking a step further, SimSiam (Chen and He 2021) discards negative pairs and the momentum encoder, relying on identical encoders and stop-gradient mechanism to prevent output collapse. Besides, DINO (Caron et al. 2021) employs vision transformers as the foundation for self-supervised learning, implementing self-distillation without requiring any labels.

Method

Overview

In this section, we describe the architecture of our proposed Token-Level Contrastive Learning with Modality-Aware Prompting (TCL-MAP) method. As illustrated in Figure 1, the framework comprises two components: Prompt-Based Augmentation and Token-Level Contrastive Learning & Intent Recognition. The former is presented following the order of Feature Extraction, Modality-Aware Prompting, and Augmented Sample Construction while the latter elucidated from the dual perspectives of Token-Level Contrastive Learning and Intent Recognition.

Prompt-Based Augmentation

Feature Extraction For the text modality, we align with established practices in text intent recognition methods (Zhang et al. 2023a,b), using BERT (Devlin et al. 2018), a powerful pre-trained language model, to extract the features. To capitalize on the fine-tuning process for mining semantic information from augmented samples in later stages, we exclusively rely on the embedding layer to extract features from the text modality at this step. Specifically, Given an input utterance t, we get all the token representations $\mathbf{Z}_{text} = [\text{CLS}, \mathbf{z}_1, \ldots, \mathbf{z}_{l_t}] \in \mathbb{R}^{(l_t+1)\times d_t}$ from the embedding layer of BERT:

$$\mathbf{Z}_{text} = \text{BERTEmbedding}(t), \tag{1}$$

where $\mathbf{Z}$ denotes the token list, $\mathbf{z}_i$ is the i^{th} token, CLS is the vector for text classification, l_t is the sequence length of text and d_t is the embedding size.

To extract video features, we employ a representative image classification model, Swin-Transformer (Liu et al. 2021), which is pre-trained on ImageNet (Deng et al. 2009). Concretely, we begin by dividing the raw video frames and get rid of $[f_1, f_2, \ldots, f_{l_v}]$. Subsequently, we obtain video feature $\mathbf{F}_{video} \in \mathbb{R}^{l_v \times d_v}$ by processing each frame f_i through Swin-Transformer to extract the feature from the last hidden layer:

$$\mathbf{F}_{video} = \text{Swin-Transformer}([f_1, f_2, \ldots, f_{l_v}]), \tag{2}$$

where f_i denotes the i^{th} frame, l_v is the number of frames and d_v is the video feature dimension. $\mathbf{F}_{video}$ is composed by concatenating all individual frame features f_i.

Ultimately, we utilize a well-established pre-trained speech recognition model, Wav2Vec 2.0 (Baevski et al. 2020), to take the output of the last hidden layer as the audio feature $\mathbf{F}_{audio} \in \mathbb{R}^{l_a \times d_a}$ for each audio segment a:

$$\mathbf{F}_{audio} = \text{Wav2Vec 2.0}(a), \tag{3}$$

where l_a is the sequence length and d_a is the audio feature dimension.

Modality-Aware Prompting Prior researches (Zhou et al. 2022b,a; Rao et al. 2022) have empirically demonstrated the effectiveness of prompt learning with multimodal data, yet they tend to ignore the inherent correlations among different modalities. Inspired by this, we propose the Modality-Aware Prompting (MAP) module which integrates text, video and audio features comprehensively to attain a premium prompt representation. As shown in Figure 2, MAP consists of two steps: Similarity-Based Modality Alignment and Prompt Generation.

In the Similarity-Based Modality Alignment step, following (Zhou et al. 2022a; Rao et al. 2022; Graves et al. 2006),we initially employ D learnable tokens to approximate the actual text prompts, aiming at avoiding the substantial burden of prompt engineering and integrating multimodal information into the prompt. Then we employ individual standardization layers for each modality's feature, which incorporates a CTC

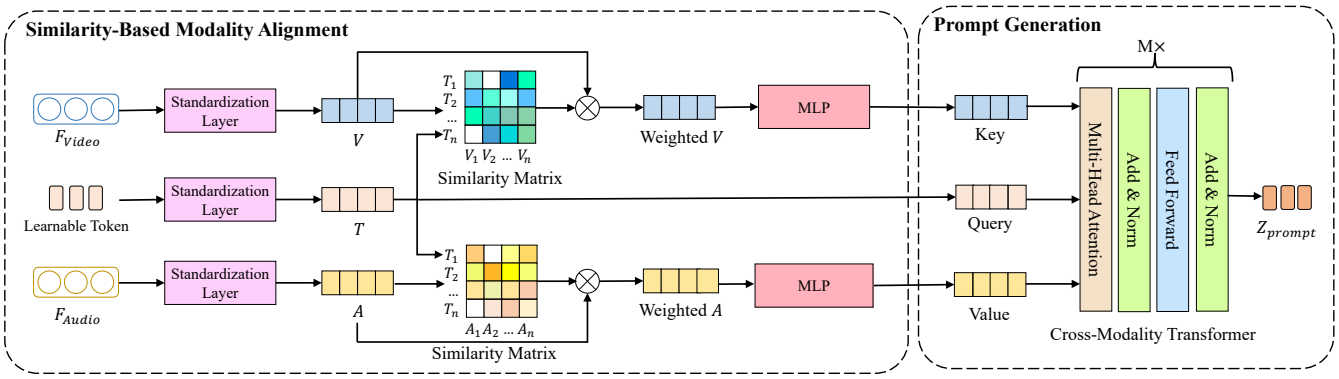

Figure 2: The details of Modality-Aware Prompting (MAP) module. We align multimodal features based on the content by computing the similarity matrix as weights and enhance correlations between modalities through a cross-modality transformer to create the modality-aware prompt.

(Graves et al. 2006) module for length normalization and an MLP to achieve consistent feature dimensions. Denoting the learnable tokens, video feature and audio feature as $\mathbf{F}_{token}$, $\mathbf{F}_{video}$ and $\mathbf{F}_{audio}$, the process can be formulated as:

$$\{\mathbf{T}, \mathbf{V}, \mathbf{A}\} = \text{MLP}(\text{CTC}(\{\mathbf{F}_{token}, \mathbf{F}_{video}, \mathbf{F}_{audio}\})), \quad (4)$$

where $\mathbf{T}, \mathbf{V}, \mathbf{A} \in \mathbb{R}^{L \times H}$ denotes the standardized features with length L and dimension H. Utilizing features within the same space, we compute two similarity matrixs $\mathbf{M}_{TV}, \mathbf{M}_{TA} \in \mathbb{R}^{L \times L}$ based on the dot product results of normalized vectors:

$$\mathbf{M}_{TV} = \alpha_{TV} \cdot \left(\frac{\mathbf{T}}{\max_i ||T_i||_2}\right)\left(\frac{\mathbf{V}^{\mathrm{T}}}{\max_i ||V_i||_2}\right), \quad (5)$$

$$\mathbf{M}_{TA} = \alpha_{TA} \cdot \left(\frac{\mathbf{T}}{\max_i ||T_i||_2}\right)\left(\frac{\mathbf{A}^{\mathrm{T}}}{\max_i ||A_i||_2}\right), \quad (6)$$

where α_{TV}, α_{TA} are threshold hyper-parameters and matrix elements $\mathbf{M}_{TV}^{(ij)}$ denotes the similarity between $\mathbf{T}_i$ and $\mathbf{T}_j$. The matrix is subsequently subjected to a softmax activation to identify crucial features, serving as weights to reduce the discrepancy between nonverbal modalities and the learnable tokens:

$$\hat{\mathbf{V}} = \text{MLP}(\text{SoftMax}(\mathbf{M}_{TV})\mathbf{V}), \quad (7)$$

$$\hat{\mathbf{A}} = \text{MLP}(\text{SoftMax}(\mathbf{M}_{TA})\mathbf{A}), \quad (8)$$

where $\hat{\mathbf{V}}, \hat{\mathbf{A}} \in \mathbb{R}^{L \times H}$ denotes the aligned vectors and we designate $\hat{\mathbf{T}} = \mathbf{T}$.

During the Prompt Generation step, we leverage cross-modality attention mechanism to effectively fuse features from three modalities into a modality-aware prompt $\mathbf{Z}_{prompt}$. Following (Tsai et al. 2019), We take $\hat{\mathbf{T}}$ as the query, $\hat{\mathbf{V}}$ as the key, and $\hat{\mathbf{A}}$ as the value for the input of the multi-head attention. The attention of the i^{th} head is calculated as follows:

$$\text{Attention}_i(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i) = \text{SoftMax}\left(\frac{\mathbf{Q}_i\mathbf{K}_i^{\mathrm{T}}}{\sqrt{d_k}}\right)\mathbf{V}_i, \quad (9)$$

where $\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i$ are projected inputs and d_k denotes the feature dimension H. The attentions of all the heads are concatenated and sequentially processed through an Add&Norm layer, a Feed Forward layer, and another Add&Norm layer to produce the ultimate outputs, which constitute the modality-aware prompts $\mathbf{Z}_{prompt}$.

Augmented Sample Pair Construction It has been shown the superiority of text modality in achieving expressive representations for intent recognition (Zhang, Xu, and Lin 2021). Inspired by this, we propose a new data augmentation method, which obtains representations of the ground truth intent labels within the semantic space provided by the text modality. Furthermore, to impose additional constraints by establishing an optimal multimodal semantic environment, we employ the modality-aware prompt to influence the label token $\mathbf{z}_{label}$. To sum up, the augmented sample $\widetilde{\mathbf{Z}}$ is formed by concatenating the original text tokens $\mathbf{Z}_{text}$, the modality-aware prompt tokens $\mathbf{Z}_{prompt}$, and the ground truth label token $\mathbf{z}_{label}$, while the normal sample $\mathbf{Z}$ replaces $\mathbf{z}_{label}$ with the [MASK] token $\mathbf{z}_{mask}$:

$$\widetilde{\mathbf{Z}} = \mathbf{Z}_{text} \oplus \mathbf{Z}_{prompt} \oplus [\mathbf{z}_{label}], \quad (10)$$

$$\mathbf{Z} = \mathbf{Z}_{text} \oplus \mathbf{Z}_{prompt} \oplus [\mathbf{z}_{mask}], \quad (11)$$

where $\oplus$ denotes the concatenation operation.

Representation Learning

Token-Level Contrastive Learning To refine the augmented sample $\widetilde{\mathbf{Z}}$, we utilize a powerful multimodal fusion layer MAG-BERT (Rahman et al. 2020) to incorporate information from nonverbal modalities. For the normal sample $\mathbf{Z}$, we sorely leverage the encoder layer of BERT (Devlin et al. 2018) to ensure the stability of textual semantics. After the refinement, we extract $\mathbf{z}_{label}$ and $\mathbf{z}_{mask}$ from their respective positions to construct the pair $(\mathbf{z}_i, \mathbf{z}_j)$ for each sample and employ the NT-Xent (Sohn 2016) loss to enhance the similarity estimation within the semantic space. This involves bringing tokens from the same pair closer while pushing apart tokens that do not belong to the same pair. Assuming N represents the batch size, we can obtain a total of 2N tokens

Methods	MIntRec				MELD-DA			
	ACC (%)	WF1 (%)	WP (%)	R (%)	ACC (%)	WF1 (%)	WP (%)	R (%)
MAG-BERT	72.65	72.16	72.53	69.28	60.63	59.36	59.80	50.01
MISA	72.29	72.38	73.48	69.24	59.98	58.52	59.28	48.75
MulT	72.52	72.31	72.85	69.24	60.36	59.01	59.44	49.93
TCL-MAP	**73.62**	**73.31**	**73.72**	**70.50**	**61.75**	**59.77**	**60.33**	**50.14**
Δ	0.97↑	0.93↑	0.24↑	1.22↑	1.12↑	0.41↑	0.53↑	0.13↑

Table 1: Multimodal intent recognition results on the MIntRec dataset and the MELD-DA dataset. Δ represents the maximum enhancement attained by our method compared to the baseline across the evaluation metrics.

Methods	Complain	Praise	Apologise	Thank	Criticize	Care	Agree	Taunt	Flaunt	Oppose	Joke
MAG-BERT	67.65	86.03	97.76	96.52	49.02	85.59	91.60	15.78	47.09	33.97	37.54
MISA	63.91	86.63	97.78	98.03	53.44	87.14	92.05	22.15	46.44	36.15	38.74
MulT	65.48	84.72	97.93	96.83	49.72	88.12	92.23	26.12	48.91	34.68	33.95
TCL-MAP	**68.70**	**87.20**	97.70	<u>97.00</u>	<u>51.30</u>	86.80	**93.10**	17.20	**50.80**	<u>35.90</u>	29.00
Human	80.08	93.44	96.15	96.90	72.21	96.09	87.21	65.55	78.10	69.04	72.22

Methods	Inform	Advise	Arrange	Introduce	Comfort	Leave	Prevent	Greet	Ask for help
MAG-BERT	71.00	69.30	63.82	67.42	76.43	75.77	85.07	91.06	64.44
MISA	70.18	69.56	67.32	67.22	78.78	77.23	83.30	82.71	67.57
MulT	70.85	69.43	65.44	71.19	76.44	75.58	81.68	86.65	69.12
TCL-MAP	**72.80**	68.90	65.40	<u>68.40</u>	**79.80**	**83.40**	<u>83.60</u>	<u>90.10</u>	66.40
Human	79.69	87.14	81.40	84.09	95.95	97.06	86.43	94.15	88.54

Table 2: F1-Score (%) comparison between baselines and our method for each class of MIntRec. For the results of our method, bold indicates the best performance while underlining indicates the second best performance within each class.

using the aforementioned approach. The contrastive loss is computed by:

$$l_{ij} = -\log \frac{\exp\left(\text{sim}\left(\mathbf{z}_i, \mathbf{z}_j\right)/\tau\right)}{\sum_{k=1}^{2N} 1_{[k\neq i]} \exp\left(\text{sim}\left(\mathbf{z}_i, \mathbf{z}_k\right)/\tau\right)}, \quad (12)$$

$$\mathcal{L}_{con} = -\frac{1}{2N}\sum_{i,j}(l_{ij} + l_{ji}), \quad (13)$$

where $1_{[k\neq i]}$ is an indicator function evaluating to 1 iff $k \neq i$, $\text{sim}(\cdot,\cdot)$ denotes the cosine similarity between two vectors and τ denotes the temperature hyper-parameter.

Classification Adopting the widely-used approach, we utilize the mean-pooling feature $\bar{\mathbf{z}}$ of the normal tokens Z for classification and use a standard entropy loss to optimize the framework:

$$\mathcal{L}_{cls} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\phi(\bar{\mathbf{z}})_{y_i})}{\sum_{j=1}^{I}\exp(\phi(\bar{\mathbf{z}})_j)}, \quad (14)$$

where N denotes the batch size, $\phi(\cdot)$ is the classifier with a linear layer. y_i is the label of the i^{th} sample, and I is the number of labels. Ultimately, The overall learning of TCL-MAP is accomplished by minimizing the following loss:

$$\mathcal{L} = \mathcal{L}_{con} + \mathcal{L}_{cls}. \quad (15)$$

Experiments

Datasets

We conduct experiments on two challenging multimodal datasets to evaluate our proposed framework.

MIntRec MIntRec (Zhang et al. 2022) is a fine-grained dataset for multimodal intent recognition with 2,224 high-quality samples with text, video and audio modalities across 20 intent categories. We follow the dataset splits consisting of 1,334 samples for training, 445 samples for validation, and 445 samples for testing.

MELD-DA MELD-DA (Saha et al. 2020) is a large-scale dataset for dialogue act classification, comprising 9988 multimodal samples. The dataset is divided into a training set of 6,991 samples, a validation set of 999 samples and a test set of 1,998 samples. All data are annotated with 12 common dialogue act labels.

Baselines

Following (Zhang et al. 2022), we use the following state-of-the-art multimodal fusion methods as the baselines: (1) MAG-BERT (Rahman et al. 2020) presents an efficient attachment for pre-trained language models to expand the capabilities to fine-tune on multimodal representations; (2) MulT (Tsai et al. 2019) utilizes directional pairwise cross-modality

Modules			MIntRec				MELD-DA			
SBMA	MAP	TCL	ACC (%)	WF1 (%)	WP (%)	R (%)	ACC (%)	WF1 (%)	WP (%)	R (%)
×	✓	✓	73.15	72.73	73.02	69.82	61.24	59.32	59.58	49.89
×	×	✓	72.67	72.31	72.81	69.78	60.40	58.69	59.78	49.63
✓	✓	×	72.13	71.80	72.50	68.80	61.26	59.54	59.97	50.09
✓	✓	✓	**73.62**	**73.31**	**73.72**	**70.50**	**61.75**	**59.77**	**60.33**	**50.14**

Table 3: Ablation experiments of modules in TCL-MAP on the MIntRec dataset and the MELD-DA dataset. SBMA stands for Similarity-Based Modality Alignment, MAP stands for Modality-Aware Prompt and TCL stands for Token-Level Contrastive Learning. With SBMA incorporated into MAP, there exist three distinct settings.

attention without explicit alignment to address interactions between multimodal sequences; (3) MISA (Hazarika, Zimmermann, and Poria 2020) captures both modality-invariant and modality-specific features from each modality and subsequently fuses them with self-attention mechanism.

Evaluation Metrics

We adopt accuracy (ACC), weighted F1-score (WF1), weighted precision (WP) and recall (R) to evaluate the model performance, which is commonly used in classification tasks. Considering the imbalance across different categories, for the second and the third metrics, we present the scores weighted by the number of samples in each class. Higher values indicate improved performance across all metrics.

Experimental Settings

For the implementation of our proposed method, we utilize bert-base-uncased and wav2vec2-base-960h from Huggingface Library (Wolf et al. 2019) to extract text and audio features and swin_b pre-trained on ImageNet1K (Deng et al. 2009) from Torchvision Library (maintainers and contributors 2016) to extract video features. The training batch size is set to 16, while the validation and test batch sizes are both set to 8. For the total loss $\mathcal{L}$, we employ AdamW (Loshchilov and Hutter 2017) to optimize the parameters.

Results

We conduct experiments on both the MIntRec dataset and the MELD-DA dataset, comparing our approach with state-of-the-art baselines. The results are presented in Table 1 with the optimal outcomes highlighted in bold, and the enhancements of our method over the top-performing baseline are indicated by Δ.

Firstly, we observe the overall performance. As indicated by the results, our approach has consistently outperformed the current state-of-the-art methods across all four metrics on both datasets, demonstrating significant advancements. Secondly, on the MIntRec dataset, our approach demonstrates enhancements of 0.97% on ACC, 0.93% on WF1 and 1.22% on R, which indicates the robust capability of our approach to effectively leverage multimodal information for the extraction and identification of intricate intents in real-world scenarios. Thirdly, on the MELD-DA dataset, our method also achieves notable improvements on both ACC and WP metrics, despite the presence of challenging "Others" label which is difficult to distinguish. This observation showcases the effectiveness of our method in recognizing ambiguous intents such as dialogue acts.

Discussion

Effectiveness of Each Module

To further analyze the individual contributions of the modules within TCL-MAP to the overall performance, we conducted the following ablation experiments and the results are illustrated in Table 3.

Similarity-Based Modality Alignment To assess the effectiveness of similarity-based modality alignment, we replace the alignment method with a CTC module (Graves et al. 2006) which aligns multimodal features solely from a temporal perspective and disregards the correlations. As indicated by the results, the performance of TCL-MAP exhibites a reduction of more than 0.50% across most metrics for both datasets. The most significant decrease, amounting to 0.75%, is observed in the WP metric for the MELD-DA dataset. These observations illustrate the effectiveness of our proposed similarity-based modality alignment in aligning multimodal features and facilitating the extraction of semantic information. Moreover, even without the presence of similarity-based modality alignment, TCL-MAP continues to achieve superior results on the MIntRec dataset, underscoring the efficacy of the other modules.

Modality-Aware Prompting In this setting, we remove modality-aware prompting module and directly concatenate the original text tokens with the [MASK]/Label token as the augmented pair. We note more substantial reductions on MIntRec, such as a 0.95% decrease on ACC and a 1.00% decrease on WF1. Meanwhile, the performance on MELD-DA experiences notable declines on ACC, WF1 and WP metrics. We attribute this to the fact that the optimal semantic environment created by our modality-aware prompting module aids in filtering out irrelevant semantics within the [MASK]/Label token, which makes the token-level contrastive learning more precise.

Token-Level Contrastive Learning In the absence of token-level contrastive learning, we exclude the contrastive learning loss $\mathcal{L}_{con}$ from the total loss $\mathcal{L}$ and proceed with the learning process guided by classification. In this experimental setup, all of the four metrics decrease by 1.49%, 1.51%,

1.22% and 1.70% on MIntRec and the ACC metric and the WP metric decrease by 0.49% and 0.36% on MELD-DA, indicating a significant decline on performance. The experimental results suggest that our introduced token-level contrastive learning effectively leverages the rich semantic information within the ground truth labels to guide the learning process of nonverbal modalities and simultaneously optimizes feature representations together with the classification guidance, leading to improved performance.

Performance of Fine-grained Intent Classes

To examine the performance of our method in each fine-grained intent classes, we compare the F1 scores of TCL-MAP and baseline methods for each intent class in MIntRec. As shown in Table 2, the results are obtained by averaging the scores over ten runs of experiments with different random seeds and for the scores of TCL-MAP we mark the best results in bold and the second best results with underlines within each class.

To begin with, we analyze the comprehensive results of our proposed TCL-MAP in comparison to the baseline methods. Remarkably, across all 20 intent categories, our approach attains top-2 scores in 13 categories, comprising 7 highest scores and 6 second highest scores, which indicates that TCL-MAP achieves better performance than the majority of baselines across various classes. Specifically, in categories like"Complain", "Agree" and "Leave", TCL-MAP consistently outperforms the best baseline by over 1%. Significantly, the "Leave" category exhibits the most substantial improvement of 7.63%. The significant gains can be attributed to TCL-MAP's utilization of modality-aware prompts for better text representation, which in turn enhances video and audio learning through token-level contrastive learning. Nevertheless, in the "Taunt" and "Joke" categories, TCL-MAP seems to provide less assistance in recognizing the intent, which could be caused by a combination of factors, including the limited availability of data within these categories and the intricate nature of the intents themselves.

On the other hand, we evaluate the efficacy of TCL-MAP in comparison to the human performance. From the results, we can observe that humans achieve the best performance in the majority of intent categories, which confirms the strong ability of humans to process multimodal information and infer intents through them. However, TCL-MAP surpasses human performance in the "Apologize," "Thank," and "Agree" classes, showcasing the stability of our method when handling challenging samples where humans may make mistakes. In addition, TCL-MAP has approached human performance in intent categories (e.g. "complain", "praise" and "care") which involve distinct emotional aspects and also achieved comparable performance to humans in intent categories (e.g. "Inform", "leave" and "prevent") which require an understanding of actions. These findings further validate the capability of TCL-MAP to effectively extract fearures related to human intents from raw multimodal data, such as expressions, tone of speech and movements.

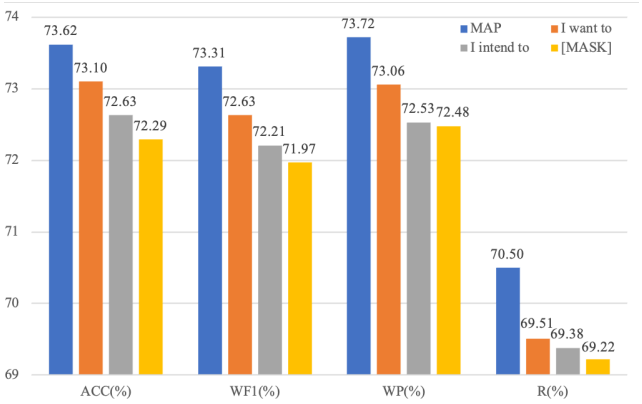

Figure 3: The comparison between Handcraft Prompt and Modality-Aware Prompt

Comparison between Handcraft Prompt and Modality-Aware Prompt

To further analyze the superiority of our modality-aware prompt, we conduct experiments with handcrafted prompt and modality-aware prompt respectively. Concretly, we select the MIntRec dataset for our experiments, driven by the fact that certain labels (e.g. "Others") in the MELD-DA dataset do not strictly represent intent categories. To make comparison, we design two handcraft prompts aimed at expressing ideas or intents, "I want to" and "I intend to", which maintain the same positions and lengths with the modality-aware prompt. Besides, we conduct an additional set of experiments using [MASK] as the prompt to demonstrate the effectiveness.

As shown in Figure 3, we observe a substantial performance advantage in the model that employs the modality-aware prompt in comparison to models using handcrafted prompts, thanks to better integration of non-textual modalities enhancing textual intent semantics extraction. Conversely, the [MASK] prompt shows a notable performance decline compared to handcrafted prompts, highlighting the risk of inappropriate prompts misleading intent understanding. Our modality-aware prompt incorporates the instance-conditional prompt concept of CoCoOp (Zhou et al. 2022a), thereby mitigating this drawback.

Conclusion

In this paper, we propose a novel Token-Level Contrastive Learning with Modality-Aware Prompting (TCL-MAP) method for multimodal intent recognition. By strengthening the correlations among modalities, our method generate the modality-aware prompt to construct an optimal multimodal semantic space for enhancing the refinement of the text modality. In return, the attained textual representation, enriched with semantics from the ground truth label token, guides the learning process of nonverbal modalities through the token-level contrastive learning. Extensive experiments on two benchmark datasets demonstrate that our approach outperforms state-of-the-art methods and carries significant implications for multimodal prompt learning.

Acknowledgements

This work is founded by the National Natural Science Foundation of China (Grant No. 62173195), National Science and Technology Major Project towards the new generation of broadband wireless mobile communication networks of Jiangxi Province (Grant No.20232ABC03402), High-level Scientific and Technological Innovation Talents "Double Hundred Plan" of Nanchang City (Grant No. Hongke Zi (2022) 321-16), and Natural Science Foundation of Hebei Province, China (Grant No. F2022208006).

References

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 12449–12460. Curran Associates, Inc.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

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

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dong, J.; Fu, J.; Zhou, P.; Li, H.; and Wang, X. 2022. Improving spoken language understanding with cross-modal contrastive learning. *Interspeech. ISCA*.

Gan, Y.; Bai, Y.; Lou, Y.; Ma, X.; Zhang, R.; Shi, N.; and Luo, L. 2023. Decorate the newcomers: Visual domain prompt for continual test time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 7595–7603.

Graves, A.; Fernández, S.; Gomez, F.; and Schmidhuber, J. 2006. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, 369–376. New York, NY, USA: Association for Computing Machinery. ISBN 1595933832.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33: 21271–21284.

Han, W.; Chen, H.; Gelbukh, A.; Zadeh, A.; Morency, L.-p.; and Poria, S. 2021. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, 6–15.

Han, W.; Chen, H.; and Poria, S. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *arXiv preprint arXiv:2109.00412*.

Hazarika, D.; Zimmermann, R.; and Poria, S. 2020. MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, 1122–1131. New York, NY, USA: Association for Computing Machinery. ISBN 9781450379885.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

Hou, M.; Tang, J.; Zhang, J.; Kong, W.; and Zhao, Q. 2019. Deep multimodal multilinear fusion with high-order polynomial pooling. *Advances in Neural Information Processing Systems*, 32.

Li, D.; Li, J.; Li, H.; Niebles, J. C.; and Hoi, S. C. 2022. Align and Prompt: Video-and-Language Pre-Training With Entity Prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4953–4963.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022.

Liu, Z.; Shen, Y.; Lakshminarasimhan, V. B.; Liang, P. P.; Zadeh, A.; and Morency, L.-P. 2018. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

maintainers, T.; and contributors. 2016. TorchVision: PyTorch's Computer Vision library. https://github.com/pytorch/vision.

Paraskevopoulos, G.; Georgiou, E.; and Potamianos, A. 2022. Mmlatch: Bottom-up top-down fusion for multimodal sentiment analysis. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4573–4577. IEEE.

Rahman, W.; Hasan, M. K.; Lee, S.; Zadeh, A.; Mao, C.; Morency, L.-P.; and Hoque, E. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, 2359. NIH Public Access.

Rao, Y.; Zhao, W.; Chen, G.; Tang, Y.; Zhu, Z.; Huang, G.; Zhou, J.; and Lu, J. 2022. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18082–18091.

Saha, T.; Patra, A.; Saha, S.; and Bhattacharyya, P. 2020. Towards Emotion-aided Multi-modal Dialogue Act Classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4361–4372. Online: Association for Computational Linguistics.

Sohn, K. 2016. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29.

Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, 776–794. Springer.

Tsai, Y.-H. H.; Bai, S.; Liang, P. P.; Kolter, J. Z.; Morency, L.-P.; and Salakhutdinov, R. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, 6558. NIH Public Access.

Wang, Z.; Zhang, Z.; Lee, C.-Y.; Zhang, H.; Sun, R.; Ren, X.; Su, G.; Perot, V.; Dy, J.; and Pfister, T. 2022. Learning To Prompt for Continual Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 139–149.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3733–3742.

Ye, M.; Zhang, X.; Yuen, P. C.; and Chang, S.-F. 2019. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6210–6219.

Yu, T.; Gao, H.; Lin, T.-E.; Yang, M.; Wu, Y.; Ma, W.; Wang, C.; Huang, F.; and Li, Y. 2023. Speech-Text Pre-training for Spoken Dialog Understanding with Explicit Cross-Modal Alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7900–7913. Toronto, Canada: Association for Computational Linguistics.

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Zadeh, A.; Liang, P. P.; Mazumder, N.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Zhang, H.; Li, X.; Xu, H.; Zhang, P.; Zhao, K.; and Gao, K. 2021a. TEXTOIR: An Integrated and Visualized Platform for Text Open Intent Recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, 167–174.

Zhang, H.; Xu, H.; and Lin, T.-E. 2021. Deep open intent classification with adaptive decision boundary. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 14374–14382.

Zhang, H.; Xu, H.; Lin, T.-E.; and Lyu, R. 2021b. Discovering new intents with deep aligned clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 14365–14373.

Zhang, H.; Xu, H.; Wang, X.; Long, F.; and Gao, K. 2023a. A Clustering Framework for Unsupervised and Semi-supervised New Intent Discovery. *IEEE Transactions on Knowledge and Data Engineering*, 1–14.

Zhang, H.; Xu, H.; Wang, X.; Zhou, Q.; Zhao, S.; and Teng, J. 2022. MIntRec: A New Dataset for Multimodal Intent Recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, 1688–1697. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392037.

Zhang, H.; Xu, H.; Zhao, S.; and Zhou, Q. 2023b. Learning Discriminative Representations and Decision Boundaries for Open Intent Detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 1611–1623.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.