Appendix

Anonymous submission

Method

Dynamic Multimodal Fusion Algorithm

This section describes the overall CL-DMDF algorithm. Given a multimodal dataset D, a temperature parameter τ, and the number of iterations num, the model initializes inputs and outputs the fused features. The detailed computation is shown below.

Algorithm 1: CL-DMDF Algorithm

Input: Multimodal dataset $\mathcal{D}$, hyperparameter τ, iteration count num

Output: Fused feature F_{fused}

Initialize $M \leftarrow \text{uniform}(-\frac{1}{n}, \frac{1}{n})$ for $M \in \mathcal{D}$
$E_m \leftarrow \text{uniform}(-\frac{1}{n}, \frac{1}{n})$ for $m \in \mathcal{M}$
Apply encoding to extract embeddings λ
for $1 < M < \mathcal{M}$ **do**
  $F_m \leftarrow \text{uniform}(-\frac{1}{n}, \frac{1}{n})$ for $m \in \mathcal{M}$
  $E_m \leftarrow \text{computer\_embeddings}(F_m, \lambda)$ for $m \in \mathcal{M}$
end for
loop
while $L < num$ **do**
  for $m \in \mathcal{M}$ **do**
    $f_m \leftarrow extract\_features(M)$
  end for
  for $i \in G^+$ **do**
    $U_{B_i} = \{F_m^1, F_m^2, ..., F_m^k\}$
    $S_i \leftarrow U_{B_i} C_i - C_i$
    $q_i \leftarrow \frac{\exp(F_i^q \cdot E_m)}{\sum_t \exp(F_t^c \cdot E_m)}$ {similarity calculation}
    $\mathcal{L}_{contra} = -\frac{1}{N} \sum \log \frac{\exp(F_i^q \cdot F_i^c)}{\sum_t \exp(F_t^c \cdot F_i^q) + q_i}$
    $update\_embeddings(E_m, \mathcal{L}_{contra})$
  end for
  for $m \in \mathcal{M}$ **do**
    $N = \frac{1}{W} \sum_{w=1}^{W} F_w^c$
  end for
  $\vec{\gamma} = \frac{N_i \sum_{E_m}}{N_i + N_c(\sum F_m^c + M)}$ {attention weight}
  $F_{fused} = \sum_{m \in \mathcal{M}} (\vec{\gamma}_i \cdot E_m)$ {fused output}
end while
end loop

- **Step 1**: Initialize all modalities and their corresponding feature embeddings, then extract and embed modality-specific features from the dataset.
- **Step 2**: Extract modality features and enhance embeddings by minimizing a temperature-scaled contrastive loss between positive and negative pairs.
- **Step 3**: Compute attention scores based on feature and modality averages, and fuse features via the adaptive module to obtain the final representation.

Experiment

Evaluation Indicators

When evaluating on the MM-IMDB dataset, we adopt both Micro-F1 and Macro-F1 metrics. Micro-F1 computes the overall F1-score by aggregating global true positives (TP), false positives (FP), and false negatives (FN) across all classes. It first calculates global precision and recall, and then derives the final F1-score. This metric is more appropriate for class-imbalanced datasets, emphasizing the model's overall performance. In contrast, Macro-F1 calculates the F1-score for each class individually and then averages them, treating all classes equally regardless of their frequency. This makes it suitable for balanced datasets or when class-wise performance is of interest. The formulas are defined as follows in Equations 1–6:

$$P_{\text{Micro}} = \frac{\sum_{c \in \mathcal{C}} TP_c}{\sum_{c \in \mathcal{C}} (TP_c + FP_c)} \tag{1}$$

$$R_{\text{Micro}} = \frac{\sum_{c \in \mathcal{C}} TP_c}{\sum_{c \in \mathcal{C}} (TP_c + FN_c)} \tag{2}$$

$$F1_{\text{Micro}} = \frac{2 \cdot P_{\text{Micro}} \cdot R_{\text{Micro}}}{P_{\text{Micro}} + R_{\text{Micro}}} \tag{3}$$

$$P_{\text{Macro}} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{TP_c}{TP_c + FP_c} \tag{4}$$

$$R_{\text{Macro}} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{TP_c}{TP_c + FN_c} \tag{5}$$

$$F1_{\text{Macro}} = \frac{2 \cdot P_{\text{Macro}} \cdot R_{\text{Macro}}}{P_{\text{Macro}} + R_{\text{Macro}}} \tag{6}$$

Among them, P denotes precision, R denotes recall, and $F1$ denotes the harmonic mean of precision and recall.

Micro-F1 computes the overall $F1$ by aggregating the predictions across all classes, while Macro-F1 calculates the $F1$ score for each class individually and then averages them.

For the NYU Depth V2 dataset, we adopt Mean Intersection over Union (Miou) as the evaluation metric. Miou is widely used in segmentation tasks, especially for assessing class-wise partition accuracy. It effectively quantifies model performance across all classes by computing the intersection over union (IoU) between predicted and ground-truth regions. The average IoU across all classes is defined as:

$$\text{MIoU} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{TP_c}{TP_c + FP_c + FN_c} \tag{7}$$

For the CMU-MOSEI dataset, we use Accuracy (Acc) and Mean Absolute Error (MAE) as evaluation metrics. Accuracy measures the proportion of correctly predicted samples, reflecting the model's overall classification performance:

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

MAE quantifies the average prediction error, defined as:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{9}$$

Lower MAE values indicate smaller deviations between predictions and ground truth, implying higher model efficiency.