Supplementary Materials for Efficient Multimodal Fusion via Interactive Prompting

Yaowei Li [1] Ruijie Quan [2] Linchao Zhu [2] Yi Yang [2]

yaowei.li@uts.edu.au, {quanruijie, zhulinchao, yangyics}@zju.edu.cn

[1] ReLER, AAII, University of Technology Sydney
[2] ReLER, CCAI, Zhejiang University

1. Implementation details

1.1. Inputs and Augmentation

For images, we use a combination of 256 $\times$ 256 resize, 224 $\times$ 224 centercrop, and a random horizontal flip for augmentation and extract 16 $\times$ 16 patches per image. For texts, we tokenize the raw texts with a WordPiece tokenizer. The maximum length of the text sequences is 512 for MM-IMDB and UPMC Food-101 datasets, and 70 for the SNLI-VE dataset.

1.2. Network training

We have two different training settings for the main results in Table 2 and ablation studies in Section 4.5.

In the main results, we sweep the learning rate in $\{0.1, 0.01\}$ and early-stop on validation accuracy for UPMC Food-101 and SNLI-VE datasets, and F1-Macro for MM-IMDB. the learning rate decreases by 0.1 after three non-improvement epochs and training ends after seven non-improvement epochs. This ensures fairness as different models require varying training epochs.

All experiments in the ablation studies are trained for 30 epochs with the same cosine-decayed learning rate starting from 0.01. We use the same learning rate schedule and a fixed number of training epochs as we only need to train PMF in ablation study.

1.3. PMF with NAS

Search Space. The search space contains fusion layer $L_f \in \{L-6, L-4, L-2\}$ and prompt length $M \in \{2, 4, 8, 16\}$. Specifically, the length of any prompt vector used in PMF is chosen from $\{2, 4, 8, 16\}$. Before each iteration in the training stage, we first sample the fusion layer L_f from $\{L-6, L-4, L-2\}$, and then we randomly sample the length for every prompt vector. The weight entanglement strategy allows the weight-sharing of the same prompt vector between different samplings. For example, when M^l_{qcp} is sampled 4 and 16 in two consecutive iterations, the first four prompts in $\mathbf{z}^l_{qcp}$ will be updated twice in both iterations while the last twelve prompts in $\mathbf{z}^l_{qcp}$ will only be updated in the second iteration. All sampling processes in the training follow the uniform probability.

Evolution Search. In the searching stage, we first randomly pick 30 candidate sets as seeds and find top-10 sets to generate the next generation of architectures through mutation and crossover. For mutation, a candidate set first mutates the fusion layer with a probability of 0.2 and then mutates every prompt length to a random length in $\{2, 4, 8, 16\}$ with a probability of 0.2. For a crossover, we randomly select two candidate sets and use the hyper-parameters in these two sets to produce a new candidate set. We generate 30 new candidates per searching epoch for 5 epochs in total.

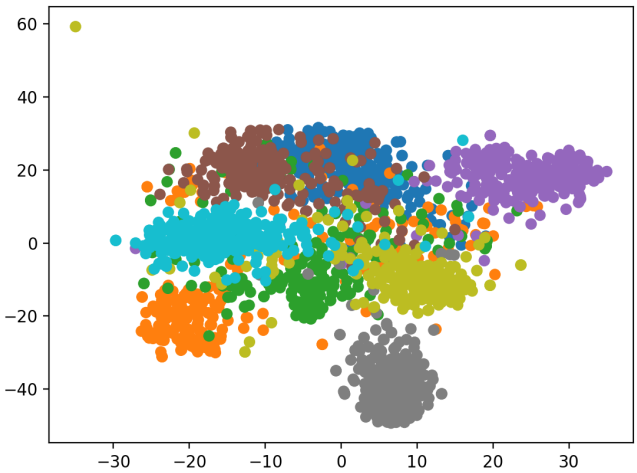
2. Visualization

To understand the effect of our proposed prompt-based multimodal fusion method, we visualize the output feature of the CLS tokens of different multimodal fusion layers via t-SNE (*i.e.* $z^l_{CLS-img}$ and $z^l_{CLS-txt}$, where $l \geq L_f$). For a qualitative comparison, we also visualize the output feature distribution without the information from the other modality. The visualization results are shown in the figures in this supplementary material. This controlled comparison clearly demonstrates the positive impact brought by the cross-modal information fusion.

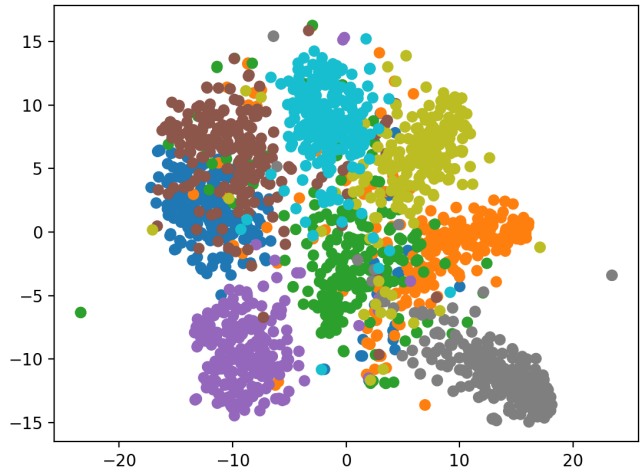
For UPMC-Food 101, we randomly sampled 8 out of 101 classes in the test set for visualization. The results shown in Fig. 1 and Fig. 2 indicates that both vision and text encoders benefit from the cross-modal information, resulting in a more desirable feature space where the samples of the same class are closer to each other compared to samples of different classes.

For visualization on SNLI-VE, we randomly sampled

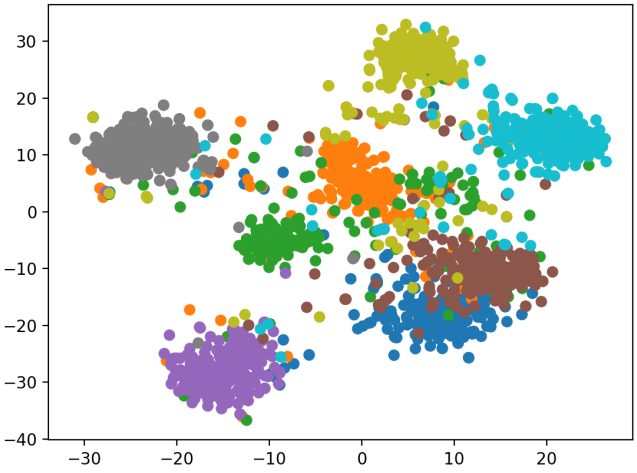
500 image-text pairs for each of the three classes in the test set. Since each image premise has three paired text hypotheses in SNLI-VE, it is meaningless to use extracted visual feature alone in the classification. The visualizations in Fig. 3c andFig. 3a shows that the visual features of three classes are not dividable without cross-modal fusion. After fusing the textual information into the vision encoder, the feature space starts to disentangle as shown in Fig. 3b and Fig. 3d. For language encoder fusing cross-modal information from vision modality, there is no significant difference shown in the visualization results in Fig. 4.



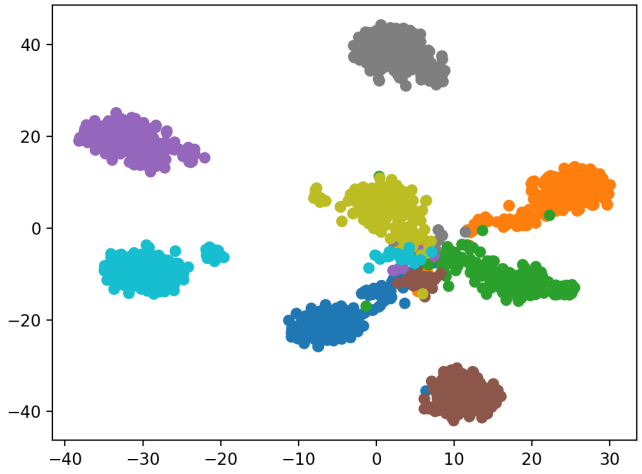
(a) Output visual feature of 11th TransLayer without cross-modal fusion.



(b) Output visual feature of 11th TransLayer fused with textual information.

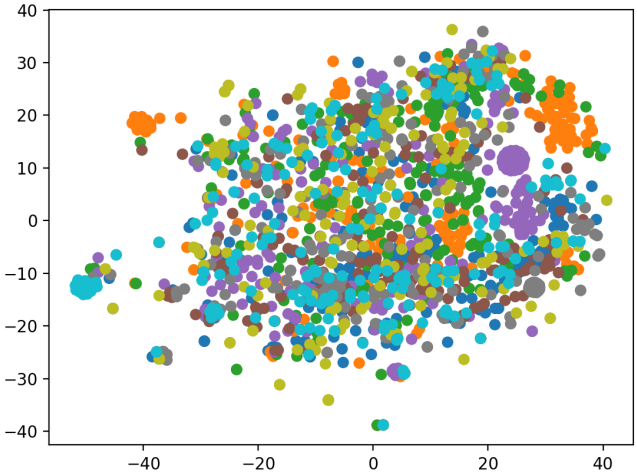


(c) Output visual feature of 12th TransLayer without cross-modal fusion.

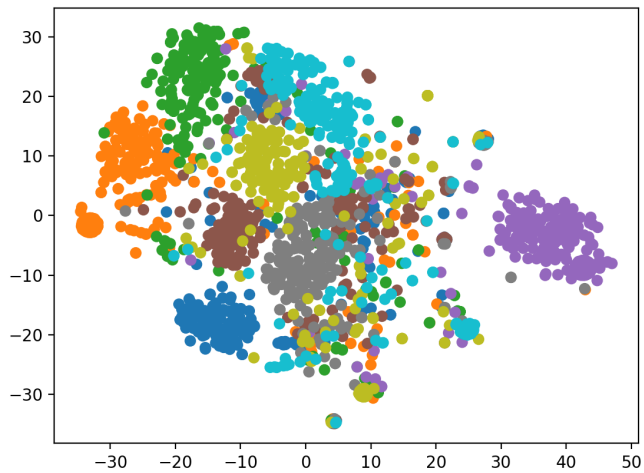


(d) Output visual feature of 12th TransLayer fused with textual information.

Figure 1. t-SNE visualization of visual features on UPMC-Food 101 test set

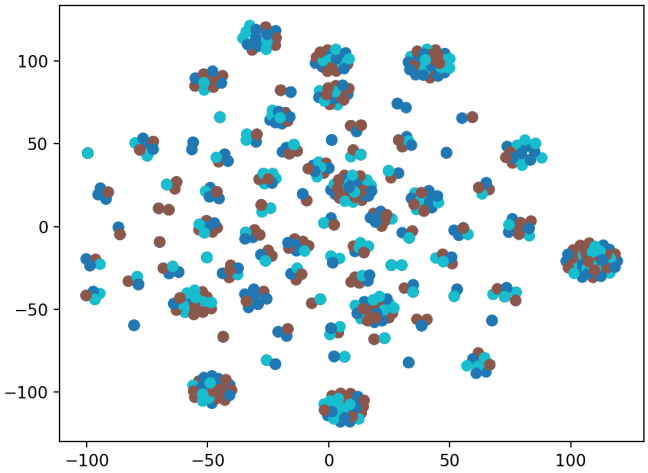


(a) Output textual feature of 12th TransLayer without cross-modal fusion.

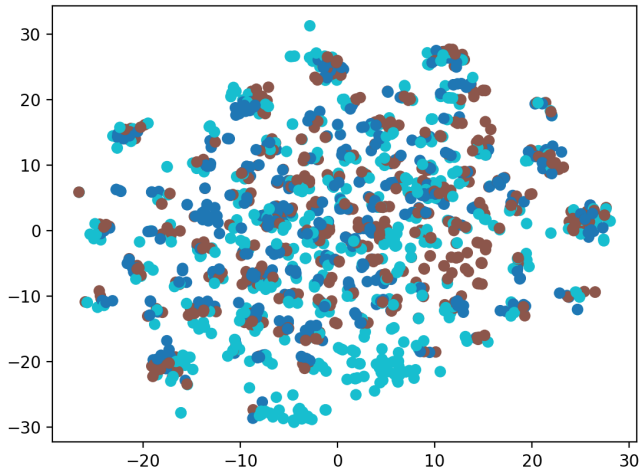


(b) Output textual feature of 12th TransLayer fused with visual information.

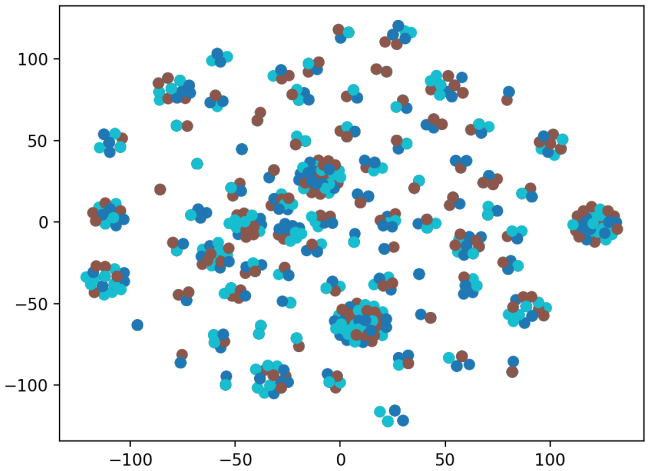
Figure 2. t-SNE visualization of textual features on UPMC-Food 101 test set



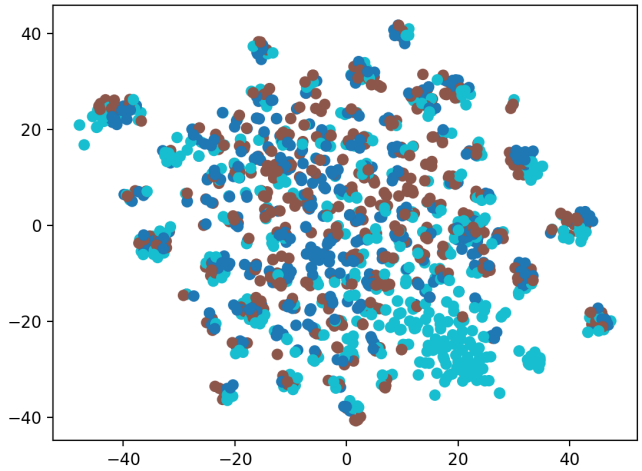
(a) Output visual feature of 11th TransLayer without cross-modal fusion.



(b) Output visual feature of 11th TransLayer fused with textual information.

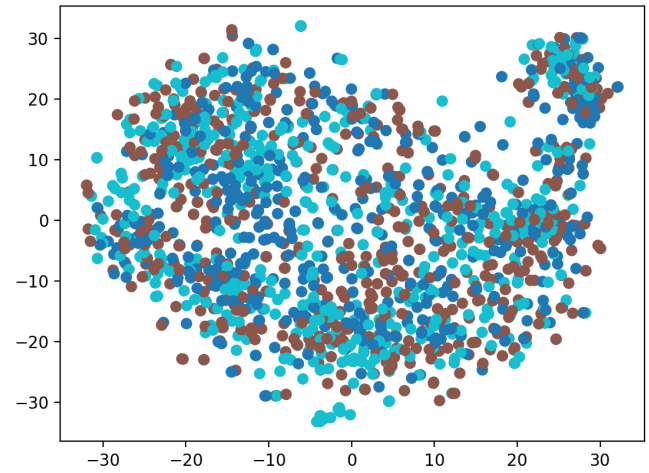


(c) Output visual feature of 12th TransLayer without cross-modal fusion.

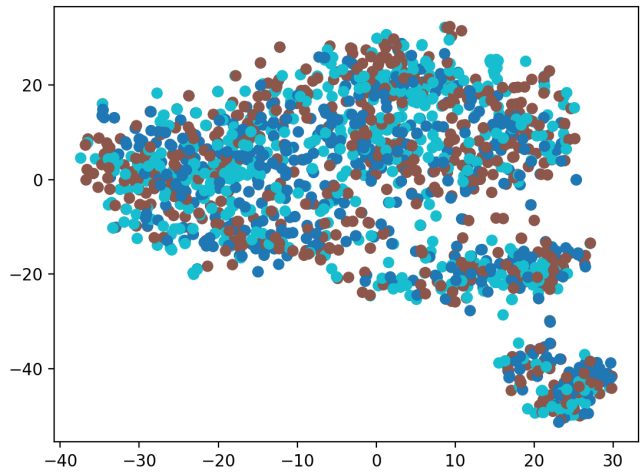


(d) Output visual feature of 12th TransLayer fused with textual information.

Figure 3. t-SNE visualization of visual features on SNLI-VE test set.



(a) Output textual feature of 12th TransLayer without cross-modal fusion.



(b) Output textual feature of 12th TransLayer fused with visual information.

Figure 4. t-SNE visualization of textual features on SNLI-VE test set