

Supplementary Materials: The Name of the Title is Hope

Anonymous Authors

1 METHODOLOGY

1.1 Keyframe-aware Masking

Given that there is a large amount of frames in a video clip, which contains redundant information. The frame sequence \bar{x}_v of a video clip contains rich priors, which explicitly correspond to neighboring frames. We can easily obtain the motion of the video frame sequence to guide the masking of redundant frames according to the temporal difference. Temporal neighbor frames in a video clip can be divided into global neighbor frames and local neighbor frames. The local and global difference information are defined as:

$$M_i^{local} = \frac{1}{2k} \left(\sum_{j=i-k}^i \bar{x}_{v_j} + \sum_{j=i+1}^{i+k} \bar{x}_{v_j} \right) - \bar{x}_{v_i} \quad (1)$$

$$M^{global} = \text{MultiHead}(\bar{x}_v, \bar{x}_v, \bar{x}_v), \quad (2)$$

where the subscript i denotes the current frame. The stride k controls the window size of the local neighbor frame. For both ends of the video frame sequence, we employ replicate padding strategy [2] to pad the original sequence length T_v to target sequence length $T_v + 2k$. The first frame is repeated k times for the beginning and the last frame is repeated k times for the end. For global difference information, we utilize multi-head attention [5] to capture the relative dependencies of all frames. The local-global embeddings $M = [M^{local}, M^{global}]$ passes through a Multi-Layer Perceptron (MLP) to predict the probability whether to mask the video frame. Formally,

$$\pi = \text{Softmax}(\text{MLP}(M)), \pi \in \mathbb{R}^{T_v \times 2}, \quad (3)$$

where the probability of index '0' ($\pi_{i,0}$) of π means to mask this video frame, and the probability of index '1' ($\pi_{i,1}$) means to keep this video frame. The subscripts i represents i -th frame in the video clip. We can easily obtain the keyframe masking decision vector D by sampling from probability π and drop the uninformative frame $x_v = \bar{x}_v \odot D$ [4]. To ensure that the sparse video frame sequence \hat{x}_v and the original sequence x_v have similar semantics in the embedding space, we employ *gated recurrent units* GRU [1] and L2 regularization to compute video frame sequence reconstruction loss:

$$\mathcal{L}_{recon} = \|\text{GRU}(x_v) - \text{GRU}(\hat{x}_v)\|_2 \quad (4)$$

2 EXPERIMENTS SETTING

2.1 Datasets

CMU-MOSI [6]. This dataset is consist of 2199 videos, which contains manually transcribed text, audio and visual modal information. The training set, validation set, and test set each contained 1284, 229, and 686 samples. The label is an sentiment score (on a range of -3 to 3). Where sentiment score greater than 0 is positive, less than 0 is negative, and equal to 0 is neutral.

CMU-MOSEI [7]. The dataset collects of 22,856 videos from youtube, The dataset includes training dataset (16326 samples), the valid dataset (1871 samples) and the test dataset (4659 samples). The meaning of the label is the same as that of CMU-MOSI.

MELD [3]. It incorporates the same dialogues as EmotionLines, but introduces additional audio and visual modalities alongside text. Comprising over 1400 dialogues and 13000 utterances from the Friends TV series, MELD involves multiple speakers engaging in the dialogues. MELD provides sentiment annotations (positive, negative, and neutral) for each utterance. We utilize the multi-modal sentiment analysis datasets CMU-MOSI, CMU-MOSEI, and MELD to construct our training and testing sets. We train the model on the source domain and perform inference on both the source and target domains. We select to report the test set performance corresponding to the best performance observed on the validation set with 200 epochs. For datasets CMU-MOSI and CMU-MOSEI, we discretize the labels to obtain a three-class classification task. The distribution of labels (Negative, Neutral, Positive) in the three test sets are as follows : CMU-MOSI:{347, 106, 233}, MELD:{1015, 1891, 1685}, and CMU-MOSEI:{831, 1256, 521}. The three-class dataset exhibits approximate balance, and we report 3-class accuracy as the evaluation metric.

REFERENCES

- [1] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning*, December 2014.
- [2] Yunyao Mao, Jiajun Deng, Wengang Zhou, Yao Fang, Wanli Ouyang, and Houqiang Li. 2023. Masked Motion Predictors are Strong 3D Action Representation Learners. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10181–10191.
- [3] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508* (2018).
- [4] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. 2021. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems* 34 (2021), 13937–13949.
- [5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [6] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259* (2016).
- [7] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. 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)*. 2236–2246.