

Supplementary Materials: KEBR

Anonymous Authors

A DATASETS

MOSI³ [16] is a widely utilized dataset with three modalities (i.e., text, vision, and audio) specially designed for sentiment analysis. MOSI comprises speakers' opinions on various topics, such as film, extracted from 93 YouTube videos, encompassing 2,199 discourse video clips. The MOSEI⁴ [1] dataset is a larger version of MOSI, which contains 22,856 annotated video clips on 250 different topics. Each clip in both datasets was annotated with a sentiment score ranging from -3 (strongly negative) to +3 (strongly positive). Table 5 shows the statistics for all datasets.

Table 5: Dataset statistics

| | Datasets | #Speakers | # Clips | # Train | # Valid | #Test |
|--------------|-----------|-----------|---------|---------|---------|-------|
| Pre-training | VoxCeleb1 | 1105 | 132708 | 132708 | | |
| | VoxCeleb2 | 5256 | 947726 | 947726 | | |
| Testing | MOSI | 93 | 2199 | 1284 | 229 | 686 |
| | MOSEI | 1000 | 22856 | 16326 | 1871 | 4659 |

B EXPERIMENTAL DESIGN

B.1 Feature Extraction

Text: We use BERT⁵ [9] as the encoder for KEBR text. Moreover, we investigate the impact of backbone language models of varying sizes. Specifically, for pre-training text information, we utilize the Google Cloud Speech API to acquire transcripts of video content. Then, we use the VADER 2 sentiment lexicon [14] to search for and block sentiment words.

Audio: The library librosa [4] is used to extract frame-level acoustic features. These include a 12-dimensional Constant-Q Chromatogram (CQT), 20-dimensional MelFrequency Cepstral Coefficients (MFCCs), and 1-dimensional log fundamental frequency (log F0). Extracting audio signals from videos using FFmpeg⁶ tool.

Vision: We employ the widely recognized MultiComp OpenFace2.2 toolkit [3] for extracting visual features. These features include 40-dimensional rigid and non-rigid shape parameters, 340-dimensional facial landmarks, 35-dimensional facial action units, 288-dimensional eye gaze, and 6-dimensional head pose and orientation.

B.2 Hyper-Parameters Setting

Table 6 provides a detailed hyper-parameters setting.

C BASELINES.

Table 6: Hyper-parameters Setting

| Hyper-parameter | Value |
|-----------------------------------|----------|
| d_t | 768&1024 |
| d_a | 33 |
| d_v | 709 |
| K | 4 |
| g | 0.2 |
| s | 1 |
| Batch size | 32 |
| Epoch | 200 |
| Optimizer | Adam |
| Learning rate of BERT | 5e-6 |
| Learning rate of other parameters | 1e-4 |
| Fully connected layer | 256 |

TFN. The Tensor Fusion Network (TFN) [50] encodes the three modalities with embedding sub-networks. It uses the outer product to model single-peak, double-peak, and triple-peak interactions as fusion results.

LMF: Low-rank multimodal fusion (LMF) [19] utilizes a low-rank rank tensor to improve the efficiency of multimodal fusion.

MISA. The modality-invariant and -specific representations (MISA) [12] design a multitask loss including task prediction loss, reconstruction loss, similarity loss, and difference loss to learn modality-invariant and modality-specific representations.

MAG-BERT. The Multimodal Adaptation Gate for Bert [29] builds an alignment gate that allows the multimodal fusion of audio and visual information into BERT models.

HyCon. Hybrid Contrastive Learning (HyCon) [22] performs intra- and inter-modal contrastive learning in modalities to explore cross-modal dynamic interactions.

MIMM. Multimodal InfoMax (MMIM) [11] maximizes the mutual information between pairs of unimodal inputs, as well as between multimodal fusion results and unimodal inputs, to aid in the main MSA task.

ConKI. Multimodal Contrast Knowledge Injection (ConKI) [49] enhances the domain-specific knowledge representation of each modality by learning adaptor architecture-based knowledge injection along with the general knowledge representation.

MuT. Multimodal Transformer (MuT) [35] proposes directed pairwise cross-modal attention, which adapts one modality to another for multimodal fusion.

CENeT. Cross-modal Enhancement network (CENet) [37] enhances text representation by integrating visual and audio information into the language model. In addition, clustering introduces a feature transformation strategy to reduce distribution differences between verbal and non-verbal initial representations.

³ <http://multicomp.cs.cmu.edu/resources/cmu-mosi-dataset/>

⁴ <http://multicomp.cs.cmu.edu/resources/cmu-mosei-dataset/>

⁵ <https://huggingface.co/google-bert/>

⁶ <https://ffmpeg.org/>

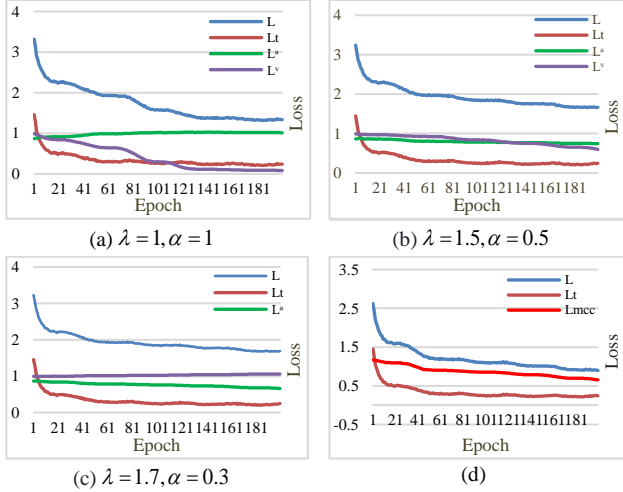


Figure 5: Modal imbalance in MSA. L , L_t , L^a , L^v and L_{mcc} represent the total loss, sentiment prediction loss (L_{task}), unimodal audio similarity loss, unimodal visual similarity constraint loss, and multimodal cosine constraint loss in this paper, respectively.

Self-MM. The Self-Supervised Multitask Learning (Self-MM) [48], proposes a label generation module based on self-supervised learning to obtain unimodal supervision. Then they joint-train the multimodal and unimodal tasks for better fusion results.

D FURTHER EXPERIMENTAL ANALYSIS

D.1 Modal Imbalance in MSA

Some studies have interpreted imbalanced optimization as a barrier to multimodal learning [40]. It shows that the problem of modal imbalance is common in multimodal tasks. To explore modal imbalance in MSA, we redesigned the comparative experimental analysis of KEBR, drawing from prior studies that addressed modal imbalance by adjusting learning rates [38,26].

In Fig. 5(a)(b)(c), the loss function is $L = L_{task} + \lambda L^a + \alpha L^v$. While keeping other calculations constant, the joint multimodal cosine constraint loss for audio and vision in Eq. (17) $L = L_{task} + L_{mcc}$ is divided into two separate cosine similarity losses. Hyperparameters λ, α are then adjusted for different losses to modify the learning rate. Fig. 5(d) uses the joint loss function in this paper, i.e., Eq. 17. In Fig. 5(a), it is observed that audio converges faster at the same learning rate, while visual loss increases as audio convergence. This suggests the existence of modal imbalance in MSA tasks. In Fig. 5(b), we apply a higher learning rate for visual loss and a lower for audio. We discovered that this merely postponed bias, as it still gradually led to an imbalance problem as the training epoch progressed. Continuously increasing the adjusted learning rate, as shown in Fig. 5(c), audio and vision still converge in the opposite trend. It shows the imbalance problem of unimodal encoding in

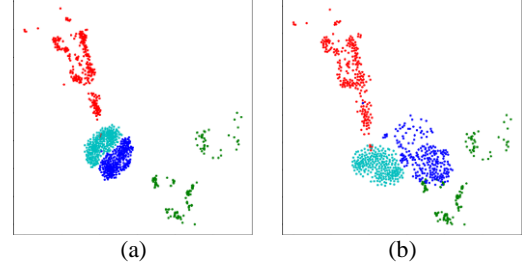


Figure 6: Visualization analysis. Red, cyan, green, and blue respectively represent audio, text, visual, and cross-modal fusion features. (a) Indicates no multimodal cosine constraint. (b) Indicates supervised multimodal cosine constraint. In (a), the text encoder corresponds to the original BERT, while in (b), it represents the KEBR pre-trained with BERT as the backbone model.

joint learning. If a non-verbal modality holds a relative advantage in fusion, it will quickly amplify this advantage. This imbalance problem is difficult to solve by controlling the learning rate on the three-modalities joint task of MSA. Compared to our proposed joint multimodal cosine constrained loss (MCC), as shown in Fig. 5(d), the loss of MCC converges stepwise with training.

D.2 Visualization Analysis

The 2D projection of modal features enables a straightforward examination of their distribution. We applied t-SNE on the verification set of MOSI to visualize the disparity in the distribution of KEBR features post multimodal fusion, both with and without MCC. Closer distributions indicate higher feature similarity.

In Fig. 6(a)(b), the text feature (cyan) is more centered than the audio (green) and visual (red), indicating the correctness of the text-based in the MAS task. At the same time, it is observed that audio and visual are distributed on either side of the fused features (blue). Without constraints, once the fused features tend to a certain non-verbal modality, the gap with the other non-verbal modality will inevitably increase relatively, resulting in a situation of opposition between the two non-verbal information in the fusion process. This also explains the opposing phenomenon observed in Fig. 5(a, b, c) of D.1 regarding the convergence loss of audio and visual features.

Further analysis of Fig. 6(a) shows that in the absence of MCC, the fused modal features (blue) are highly concentrated near the text features, but away from the audio and visual. Compared with Fig. 6(b) following the inclusion of MCC, the fused modal features are close to the text while further expanding the distribution to be close to the audio and visual feature distributions, so that the fused modal can take into account the features of non-verbal information.

Further, we observe that the text features in Fig. 6(b) exhibit greater dispersion compared to those in Fig. 6(a). The overall text representation in Fig. 6(b) is closer to audio and visual than Fig. 6(a). The overall text representation in Fig. 6(b) demonstrates a closer affinity with audio and visual modalities compared to Fig. 6(a). This suggests that the original BERT exhibits a degree of multimodal representation capability following KEBR pre-training with a substantial corpus of speaker videos. The pre-trained KEBR

KEBR: Knowledge Enhanced Self-Supervised Balanced Representation for Multimodal Sentiment Analysis

tends to encode text in a way that favors audio-visual fusion, which will facilitate representation in multimodal tasks.