Frame-level Nonverbal Feature Enhancement Based Sentiment Analysis

Cangzhi Zheng[a], Junjie Peng[a,b,*], Lan Wang[a], Li'an Zhu[a], Jiatao Guo[a] and Zesu Cai[c]

[a] *School of Computer Engineering and Science, Shanghai University, Shanghai, China*
[b] *Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai, China*
[c] *School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China*

ARTICLE INFO

Keywords:
Multimodal Sentiment Analysis
Frame-level Enhancement
Pre-trained Language Models
Vector Quantization

ABSTRACT

Multimodal Sentiment Analysis (MSA) comprehensively utilizing data from multiple modalities to obtain more accurate sentiment attribute, has important applications in other fields, such as social media analysis, user experience evaluation and medical health, etc. It is worth noting that previous studies have paid little attention to the inconsistency of the initial representation granularity between verbal (textual) and nonverbal (acoustic and visual) modalities. As a result, the imbalanced emotional information between them complicates the interaction and fusion process, and ultimately affects the model's performance. To solve this problem, this paper proposes a Frame-level Nonverbal feature Enhancement Network (FNENet) to improve performance on MSA by reducing the gap between modalities. Specifically, acoustic and visual information is integrated through the vector quantization into a pre-trained language model to enhance the textual representation and improve the FNENet's accuracy in predicting sentiment. Additionally, a sequence fusion mechanism is applied to FNENet, which benefits the word-level semantic expression according to the asynchronous affective cues preserved in unaligned frame-level nonverbal features. Extensive experiments on three benchmark datasets demonstrate that FNENet significantly outperforms baseline methods. It indicates that our model has potential application on MSA.

1. Introduction

With the explosive growth of human-centric online videos, such as YouTube and Facebook, there is growing recognition of the importance of Multimodal Sentiment Analysis (MSA) in computer science academia. MSA mainly aims to perform emotion recognition and sentiment analysis through multimodal signals, such as text, acoustics, and vision (Zadeh et al., 2018; Hazarika et al., 2020). In recent years, related research has continuously emerged analyzing human sentiment in videos using language (textual), acoustic, and visual patterns, and some studies (Han et al., 2021; Wu et al., 2022; Lin et al., 2023; Peng et al., 2023; Zhao et al., 2023) have made significant progress on MSA.

Among these studies, Pre-trained Language Models (PLM) based MSA studies are the most popular ones widely studied in recent years. These models, specifically Transformer based models (Devlin et al., 2019; Yang et al., 2019; Liu et al., 2019), can extract contextual semantic features and are very flexible for downstream tasks through fine-tuning. It keeps them vastly improving the recognition accuracy on MSA (Hazarika et al., 2020; Sun et al., 2020; Yu et al., 2021). However, the significant differences between modalities caused by heterogeneity at the data level limit the ability to achieve higher performance in the fusion phase (see Fig. 1).

In these studies, the text feature pre-trained by PLM is denoted as verbal modality, while the acoustic and visual features extracted by feature extraction tools (Degottex et al., 2014; Baltrusaitis et al., 2016; McFee et al., 2015; Baltrusaitis et al., 2018) are collectively named nonverbal modalities.

In most prior studies, researchers have focused on designing more effective mechanisms for integrating modalities to improve accuracy while overlooking the heterogeneity of different modalities and the semantic-level differences that arise from using different feature extractors. For acoustic and visual modalities, hand-crafted low-level features are usually first extracted using feature extraction tools such as COVAREP (Degottex et al., 2014) and OpenFace (Baltrusaitis et al., 2016), followed by Recurrent Neural Network (RNN) based networks, such as Bidirectional Long Short-Term Memory network (BiLSTM) (Hochreiter & Schmidhuber, 1997) and Bidirectional Gate Recurrent Unit (BiGRU) (Cho et al., 2014).

In contrast, mainly studies use PLM to gain textual features. It is worth noting that nonverbal features are relatively underdeveloped compared to those verbal features learned by PLM, implying the difference in initial representation granularity between modalities (Wang et al., 2022). Thus, it makes the interaction and fusion in modalities very inefficient and eventually affects the model's performance.

As Fig. 1 shows, verbal modality (top part) features are generally high-dimensional abstract representations trained by large-scale pre-trained language models, with a granularity of strong contextual semantic correlation at the word-level. They have a high density of emotional information. On the contrary, for nonverbal modalities (bottom part), researchers typically sample the raw data frames and use feature extraction tools to get manual features, which are

*Corresponding author

cangzhizheng@shu.edu.cn (C. Zheng); jjie.peng@shu.edu.cn (J. Peng); wanglan1997@shu.edu.cn (L. Wang); blossom@shu.edu.cn (L. Zhu); guojiatao@shu.edu.cn (J. Guo); caizesu@hit.edu.cn (Z. Cai)

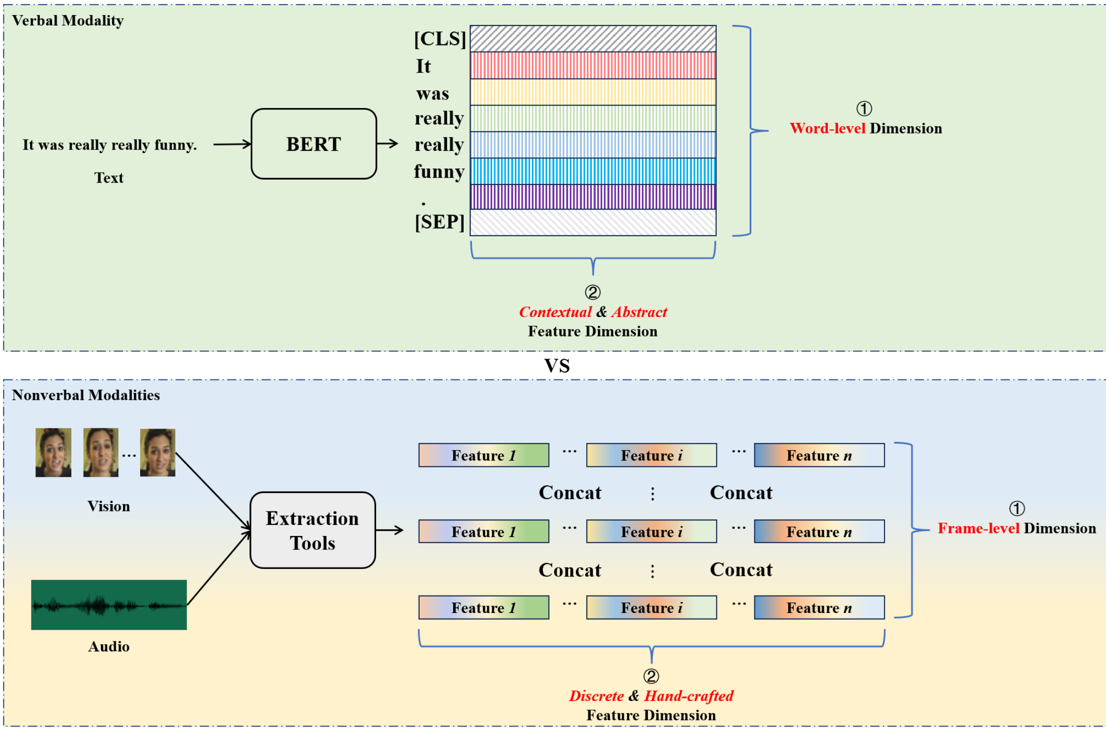


Fig. 1. Illustration of heterogeneous differences between initial feature representations of verbal modality and nonverbal modality.

commonly concatenated in different aspects related to sentiments. It is worth noting that these aspects are practically uncorrelated with each other, hence the emotional information of single frame features is badly sparse.

Additionally, an utterance-level sentiment may be different under the condition of the different nonverbal information. For example, when only using unimodal text features to judge the sentiment of "this movie is crazy", linguistic ambiguity in this sentence may lead to a large gap between the predicted sentiment and the actual sentiment. As a result, the model seems prone to bias in sentiment analysis. Given the acoustic and visual modalities, e.g., loud voice and smile, which contain rich affective information, the model predicts a positive sentence combined with asynchronous sentiment cues (Zadeh et al., 2017; Wang et al., 2020), as shown in Fig. 2.

To reduce the interaction gap between modalities and integrate sentiment cues so that improve the performance of MSA, we propose a Frame-level Nonverbal feature Enhancement Network (FNENet) to improve textual representation by integrating affective information from acoustic and visual modalities. In FNENet, Vector Quantization (VQ) is utilized to transform frame-level features by training the index embeddings of each frame of acoustic and visual raw features. The Sequence Fusion mechanism (SF) is introduced to focus on capturing asynchronous nonverbal affective context from nonverbal features. The enhanced textual representation is integrated into PLM, further improving the performance of model.

The main contributions of this paper are as follows:

- A Frame-level Nonverbal feature Enhancement Network is proposed to improve textual representation by incorporating frame-level nonverbal features into PLM.
- The frame-level feature transformation is adopted to reduce the distributional differences between the original representations in modalities by learning the index embedding of frames.
- Based on the sequence fusion mechanism, temporal information is effectively leveraged to integrate asynchronous affective cues for modalities.
- Extensive experimental results on three public datasets for MSA demonstrate that our method surpasses the baseline techniques.

2. Related Work

In this section, we introduce some related work in multimodal sentiment analysis. Next, we discuss pre-trained language models. Finally, we present some studies based on vector quantization.

2.1. *Multimodal Sentiment Analysis*

Based on the text feature extractors category, previous studies on MSA can be divided into two categories. One category involves methods that do not use PLM, and the other utilizes PLM to extract text features.

For the first category, typically, these methods utilize GloVe (Pennington et al., 2014) word embeddings followed

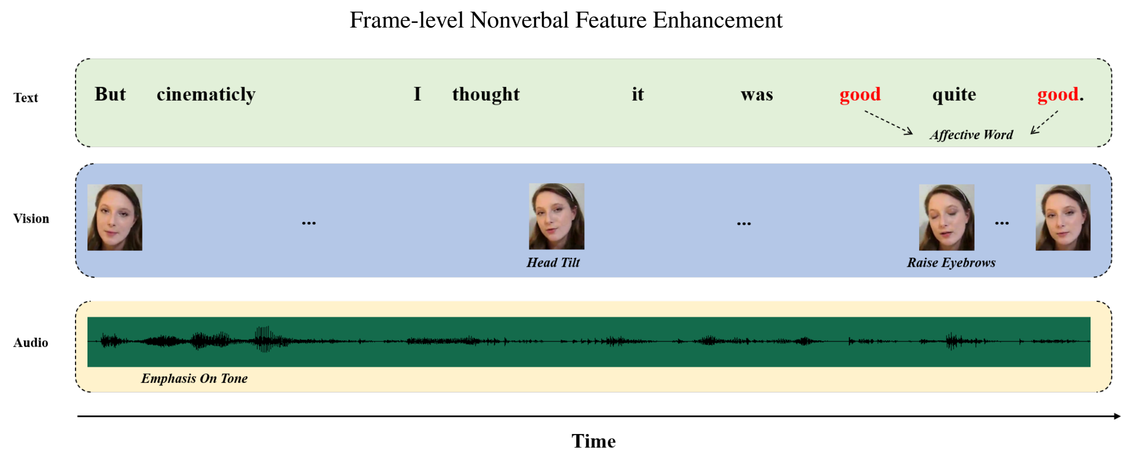


Fig. 2. It shows an example of asynchronous affective information among multiple modalities. The term "good" is considered an affective word, and the accompanying facial expressions, tone of voice, and head gestures occur at different moments during its verbalization, resulting in asynchronous affective cues across modalities.

by LSTM (Hochreiter & Schmidhuber, 1997) to extract language representations. Tensor Fusion Network (TFN) (Zadeh et al., 2017) uses a three-fold Cartesian product of three modalities to learn the intra-modal dynamics through the modality embedding sub-network. Low-rank Multimodal Fusion network (LMF) (Liu et al., 2018) reduces many parameters associated with the tensor computation by using low-rank tensors. Recurrent Attended Variation Embedding Network (RAVEN) (Wang et al., 2019) leverages fine-grained nonverbal sub-word information to dynamically adjust word representations for multimodal fusion. Factorized Multimodal Transformer (FMT) (Zadeh et al., 2019) applies Factorized Multimodal Self-attention (FMS) to design inter-modal interactions. Multimodal Transformer (MulT) (Tsai et al., 2019) uses Cross-Modal Attention (CMA), which extends the standard Transformer (Vaswani et al., 2017) model to transform one modality into another and construct the interaction between different pairs of modalities on unaligned data. On the contrary, our method focuses on asynchronous affective cues captured by the temporal attention mechanism to enhance verbal features. Further, it means the long-distance adequate information flow is unidirectionally from nonverbal to verbal.

The other category usually achieves better results than the one mentioned before because PLM trained on large text corpora can significantly facilitate the understanding of sentiment in textual modality (Wang et al., 2022). The framework of Modality-Invariant and -Specific representations for sentiment Analysis (MISA) (Hazarika et al., 2020) projects each modality into two subspaces to learn modality-invariant and modality-specific representations and fuses these two representations to predict sentiments. Interaction Canonical Correlation Network (ICCN) (Sun et al., 2020) uses canonical correlation to analyze hidden text, audio, and video relationships. Self-Supervised Multi-task Multimodal sentiment analysis network (Self-MM) (Yu et al., 2021) designs a unimodal label generation strategy to obtain unimodal labels and introduces unimodal subtasks to aid in learning modality-specific representations through a multi-task framework. Multimodal Adaptation Gate network (MAG-BERT) (Rahman et al., 2020) uses acoustic and visual features to enrich linguistic features with aligned nonverbal behavioral information, which enables BERT to adapt to multimodal inputs. Our method can capture asynchronous affective cues from unaligned nonverbal data to enhance textual representation, but MAG-BERT can only deal with the aligned data. In terms of the aligned data, Bi-Bimodal Fusion Network (BBFN) (Han et al., 2021) separates and fuses the representations of each modality to predict sentiments through an extra task loss. However, we do not employ the additional loss to enhance the model's learning on aligned data. Our method primarily focuses on sentiment prediction on unaligned data. In addition, Adaptive Multimodal Meta-Learning (AMML) (Sun et al., 2023b) introduces a meta-learning-based method to learn better unimodal representations and adapt them for subsequent multimodal fusion. Efficient Multimodal Transformer (EMT) (Sun et al., 2023a) proposes a generic and unified framework to employ utterance-level representations from each modality as the global multimodal context to interact with local unimodal features and mutually promote each other. Unlike AMML and EMT, our method pays more attention to reducing the heterogeneity between modalities to enhance the fusion effect.

2.2. *Pre-trained Language Models*

Compared with GloVe (Pennington et al., 2014), PLM has shown superior performance in textual representation. ELMo (Peters et al., 2018) has pre-trained bidirectional LSTMs on large-scale unsupervised language corpora for better performance. A standard sequence-to-sequence model is Transformer (Vaswani et al., 2017). Transformer is entirely based on the self-attention mechanism and utilizes self-attention for encoding, decoding, and information exchange between the encoder and decoder. Due to the Transformer's strong ability to represent language, large corpora containing rich language expressions (such as unlabelled data, which is easy to obtain) make it more efficient to train large-scale deep learning models. As a result, PLM can effectively represent a language's lexical, syntactic, and semantic features. Pre-trained language models, such as BERT (Devlin et al., 2019) and its variants (Yang et al., 2019; Liu et al., 2019; Brown et al., 2020), have become the core technology of current Natural Language Processing (NLP). Considering the superior performance of large pre-trained models of BERT on text (Xu et al., 2023), this paper uses the pre-trained language model BERT as the backbone network to comprehensively evaluate the FNENet framework.

2.3. Vector Quantization in Deep Learning

Vector of Local Aggregation Descriptors (VLAD) (Jégou et al., 2010) is one of VQ approaches and has tremendously impacted aggregating discriminative features for various scenarios, including video retrieval and classification. The NetVLAD (Arandjelovic et al., 2018), which extends from VLAD, is an end-to-end differentiable model that many existing neural models can easily integrate. The later NeXtVLAD (Lin et al., 2018) improves NetVLAD by significantly reducing the parameter count of the original model and improving its overall performance. The study (Wang et al., 2021) is similarly motivated to utilize NetVLAD to close the gap between learned features in text and video. Furthermore, we introduce sequential features to enhance fusion performance. The study (Hausler et al., 2021) proposes a multi-scale fusion approach by deriving patch-level features from NetVLAD residuals. Different from previous retrieval efforts, this paper draws on VQ and regards it as a discriminative feature learner to reduce the distribution difference of initial features by converting frame-level features into several cluster center embeddings. We conduct multimodal sentiment analysis research on mostly unaligned data involving text, acoustic, and visual modalities.

3. Methodology

In this section, we introduce the task setting and provide a detailed description of the proposed FNENet model.

3.1. Task Setting

Our task goal is to predict sentiment intensity variables $\hat{y} \in \mathbb{R}$ in video clips using multimodal signals. Specifically, text (t), acoustic (a), and visual (v) sequences are denoted as X_t, X_a, and $X_v \in \mathbb{R}^{S_m \times d_m}$, where $m \in \{t, a, v\}$, S_m and d_m represent the max sequence length and initial feature dimensions of modality m, respectively. The raw acoustic and visual feature sequences are denoted as $H_a \in \mathbb{R}^{S_a \times d_a}$ and $H_v \in \mathbb{R}^{S_v \times d_v}$, respectively.

3.2. Overall Model

Fig. 3 depicts the FNENet, and it is easy to notice that the model FNENet consists of two main modules: Pre-trained Language Model and Frame-level Nonverbal feature Enhancement module (FNE). The PLM serves as the backbone network for the FNENet. The FNE module comprises two sub-modules: The frame-level Feature Transformation module (FT) and the Sequence Fusion module (SF). The FT module primarily reduces the initial feature differences between text and nonverbal modalities. The SF module captures asynchronous affective cues and fuses them with verbal modality. The resulting enhanced features are embedded into the subsequent layers of the backbone network and passed through prediction layers to predict utterance-level sentiments.

3.3. Verbal Feature Extraction

To better integrate with the underlying features of the nonverbal modalities, we extract the output of a particular layer in the backbone network as the verbal feature. Specifically, the original text X_t is embedded into word vector $E_t \in \mathbb{R}^{S_t \times d_t}$ through the PLM Embedding layer, and the intermediate feature $F_t \in \mathbb{R}^{S_t \times d_t}$ is obtained after passing through the i-th layer Transformer encoder. The calculation process is formally represented as follows:

$$E_t = \text{Embedding}\left(X_t; \theta_t^{Emb}\right), \tag{1}$$

$$F_t = \text{Transformer}\left(E_t; \theta_t^{Enc}\right), \tag{2}$$

where θ_t^{Emb} and θ_t^{Enc} represent the learnable model parameters.

3.4. Frame-level Feature Transformation

Algorithm 1: Vector Quantization

Input: nonverbal frame set of $P_m \in \mathbb{R}^{N_m \times d_m}$, nonverbal query feature sequence $H_m \in \mathbb{R}^{S_m \times d_m}$, the number of clusters K_m, the maximum number of iteration M;
Output: cluster index sequence $I_m \in \mathbb{R}^{S_m}$;

```
Initialize cluster centers C_m ∈ R^{K_m×d_m}, randomly;
Initialize cluster indexes of all N_m training frames U_m ← −1;
do
    for i ← 0 to N_m − 1 do
        U_m[i] ← UpdateCenterIndex(P_m[i], C_m);
    end
    for i ← 0 to K_m − 1 do
        C_m[i] ← UpdateCenter(P_m[i], C_m);
    end
    for i ← 0 to S_m − 1 do
        I_m[i] ← QueryCenterIndex(P_m[i], C_m);
    end
while U_m reaches maximum number of iteration M or no longer changes;
return I_m
```

For pre-trained language models, the initial textual representation is a sequence of word indices in the vocabulary. However, both visual and acoustic original representations are real vector sequences concatenated by manual features. As a result of the significant initial feature difference between verbal and nonverbal modality, it is necessary to extract frame-level features for acoustics and vision and convert them into features, which are similar to that with the

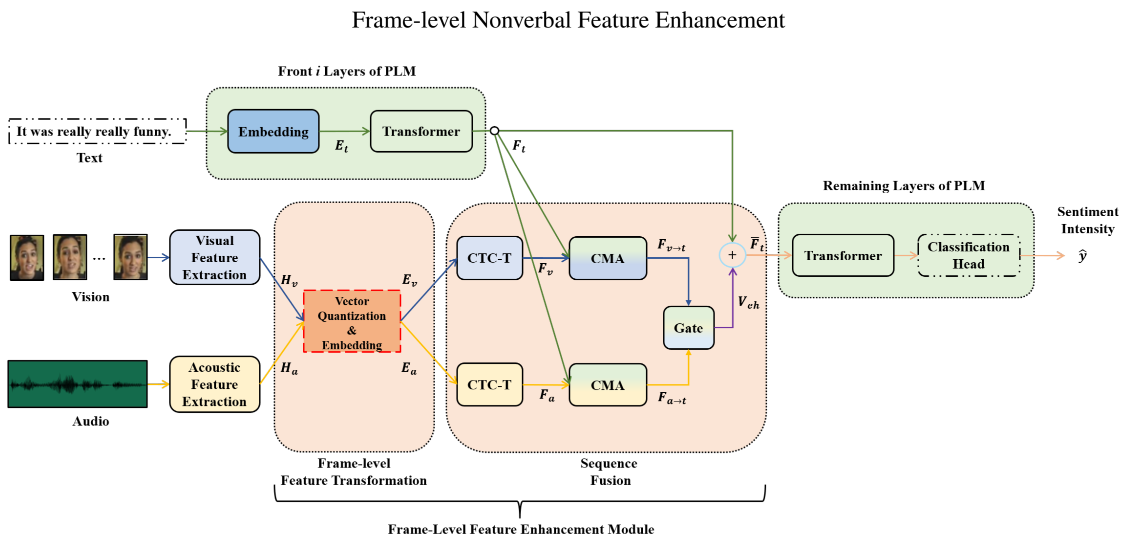


Fig. 3. FNENet comprises two components: Transformer based Pre-trained Language Model and Frame-level feature Nonverbal Enhancement module. The Frame-level feature Nonverbal Enhancement module consists of two sub-modules: The Frame-level feature Transformation module and the Sequence Fusion module.

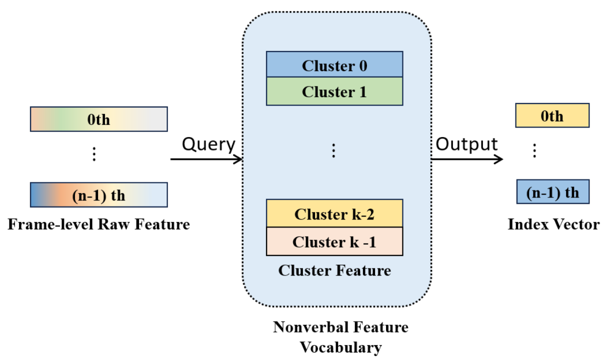


Fig. 4. Frame-level Feature Transformation on raw features. The nonverbal feature vocabulary is formed by clustering algorithm on training datasets. It is easy to transform the raw feature segments to a cluster index sequence by querying the vocabulary.

word-level granularity. The purpose is to facilitate fusion in the later stage.

In order to further narrow the distribution gap between verbal and nonverbal features, we propose a frame-level feature transformation mechanism using vector quantization, which can convert nonverbal vectors into indexes to reduce the initial distribution differences between heterogeneous patterns. Therefore, it promotes the integration of text representation and nonverbal emotional context.

Fig. 4 shows the feature transformation process of vector quantization. VQ utilizes unsupervised clustering algorithms to establish "acoustic vocabulary" and "visual vocabulary", respectively. The original feature sequence can be transformed into an index sequence by querying the nonverbal vocabulary.

Due to the low computational complexity and simplicity of the k-means method, we use k-means as the VQ to learn vocabulary from nonverbal patterns. Other clustering and dictionary learning methods can also be used to learn acoustic and visual vocabulary without losing generality.

Through VQ, we can get the query sequence I_m^i of i-th frame p_m^i from P_m. The formula is as follows:

$$C_m = \text{k-means}\left(P_m\right), \tag{3}$$

$$I_m^i = \arg\min_j \left(\|p_m^i - c_m^j\|_2\right), \tag{4}$$

$$I_m = \left\{ I_m^0, \dots, I_m^{S_m-1} \right\}, \tag{5}$$

where c_m^j is the j-th cluster center of modality m and $j \in \left[0, K_m - 1\right]$.

The obtained index sequence I_m is denoted as the representation of modality m. The VQ process is shown in Algorithm 1.

The nonverbal VQ embedding $E_m \in \mathbb{R}^{S_m \times d_m}$ is generated by VQEmbedding layer which is the similar to the composition in PLM,

$$E_m = \text{VQEmbedding}\left(I_m; \theta_m^{Emb}\right), \tag{6}$$

where θ_m^{Emb} represents the learnable model parameters of the embedding with VQ.

Due to the sparse emotional information of the original frame-level features, embedding in this way can uniformly represent frame features with similar content, and also avoid the bias caused by similar frames in different emotional orientations (Wang et al., 2022).

3.5. Sequence Fusion

Due to the influence of sampling rate, the features of audio and vision are not aligned with the text in the temporal dimension, which results in the model being unable to effectively utilize asynchronous emotional cues for more accurate judgment of emotional tendencies.

As shown in Fig. 2, facial expressions, tone of voice, and the pose of the head contain information that appears

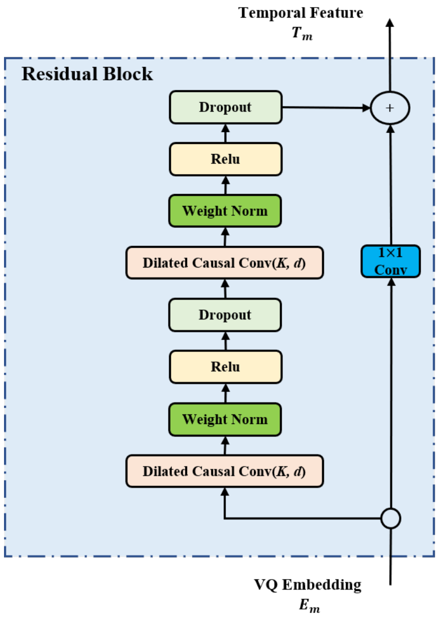


Fig. 5. Schematic diagram of a Residual Block. Each residual block contains Causal Convolution (Bai et al., 2018) with a dilated coefficient (Yu & Koltun, 2016), 1D Fully Convolutional network (FCN) (Long et al., 2015), and network optimization parts (Salimans & Kingma, 2016; Srivastava et al., 2014). The Dilated Causal Conv represents 1-D causal convolution layer with kernel size K and dilated coefficient d, while the Weight Norm operation represents weight normalization.

asynchronously when the word "good" is spoken, making it difficult to capture relevant information across different modalities (Tsai et al., 2019).

To better enhance text features, we adapt the Connectionist Temporal Classification (CTC) (Graves et al., 2006) method to align acoustic and visual modalities semantically with text modality. It is worth noting that this alignment method is pseudo alignment (in the time dimension). The purpose of the alignment is to narrow the performance gap caused by the asynchronous emotional cues of the three modalities.

In consideration that Temporal Convolutional Network (TCN) (Bai et al., 2018) has good performance advantages compared with the RNN based models do, for example, data parallelism in the processing time dimension, we use it as the sequence classifier instead of RNN based models. Additionally, TCN has a flexible receptive field and can extract local, long-distance information, which is benefit from the Dilated Convolution in the residual structure. TCN comprises multi-layer residual blocks, as shown in Fig. 5.

Our approach uses TCN based CTC module (CTC-T) to extract asynchronous contextual affective cues in nonverbal modalities. TCN outputs nonverbal temporal features $T_m \in \mathbb{R}^{S_m \times S_t}, m \in \{a, v\}$. The calculation process is as follows:

$$\mathrm{TCN}\left(E_m\right) = \mathrm{ResidualBlock}\left(E_m; \theta_m^{RB}\right), \tag{7}$$
$$T_m = \mathrm{TCN}\left(E_m\right), \tag{8}$$
$$A_m = \mathrm{Softmax}\left(T_m\right), \tag{9}$$
$$F_m' = A_m^T E_m, \tag{10}$$
$$F_m' = \mathrm{CTC}\left(E_m\right), \tag{11}$$
$$F_m = F_m' + PE_m, \tag{12}$$

where θ_m^{RB} represents the parameters in Residual Block. $A_m \in \mathbb{R}^{S_m \times S_t}$ is the attention matrix for temporal classification. $PE_m \in \mathbb{R}^{S_t \times d_m}$ is the sinusoidal position embedding (Tsai et al., 2019). $F_m \in \mathbb{R}^{S_t \times d_m}$ is final temporal feature aligned with verbal modality.

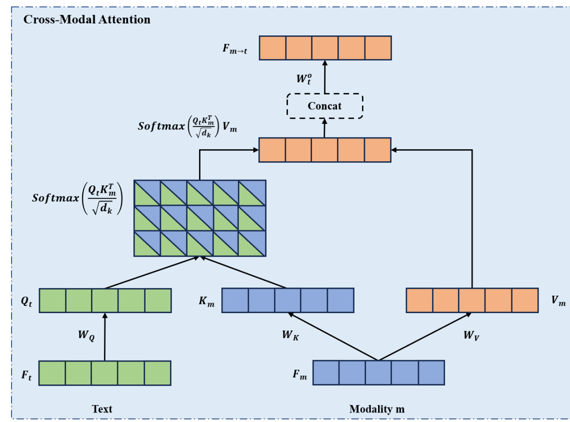


Fig. 6. Schematic diagram of the Cross-Modal Attention module. $W_Q \in \mathbb{R}^{d_t \times d_t}$ represents the query linear matrix of the text modality, while $W_K \in \mathbb{R}^{d_m \times d_t}$ and $W_V \in \mathbb{R}^{d_m \times d_t}$ represent the key linear matrix and value linear matrix of modality m, respectively. $W_t^o \in \mathbb{R}^{d_t \times d_t}$ is the mapping matrix.

The aligned temporal features F_m are processed to obtain the asynchronous, nonverbal enhanced embedding $F_{m\to t} \in \mathbb{R}^{S_t \times d_t}$ of each word from acoustics and vision through the Cross-Modal Attention (CMA) mechanism (Tsai et al., 2019). The calculation process of CMA is detailed in Fig. 6. The Querys are from the target modality t, while the Keys and Values are from the source modality m, i.e., $Q_t = F_t W_Q \in \mathbb{R}^{S_t \times d_t}$, $K_m = F_m W_K \in \mathbb{R}^{S_t \times d_t}$, $V_m = F_m W_V \in \mathbb{R}^{S_t \times d_t}$.

After splicing $F_{a\to t}$ and $F_{v\to t}$, the feature dimension is mapped back to the original feature dimension of the verbal modality through the Gate network, which is used to obtain a nonverbal enhanced embedding $V_{eh} \in \mathbb{R}^{S_t \times d_t}$ containing long-term affective context. V_{eh} is added to the original text feature F_t to obtain the embedded feature $\bar{F}_t \in \mathbb{R}^{S_t \times d_t}$. The calculation process is as follows:

$$F_{m\to t} = \mathrm{CMA}\left(F_t, F_m, F_m\right), \tag{13}$$
$$V_{eh} = \mathrm{Gate}\left(F_{a\to t}; F_{v\to t}\right), \tag{14}$$
$$\bar{F}_t = \mathrm{Avg}\left(F_t, V_{eh}\right), \tag{15}$$

where ; is concat operation, $\mathrm{Gate}(\cdot)$ is composed of fully connected dense layers, and $\mathrm{Avg}(\cdot)$ means the average function.

Fig. 7 shows an example of updating word sentiment semantic representation with nonverbal feature enhancement information.

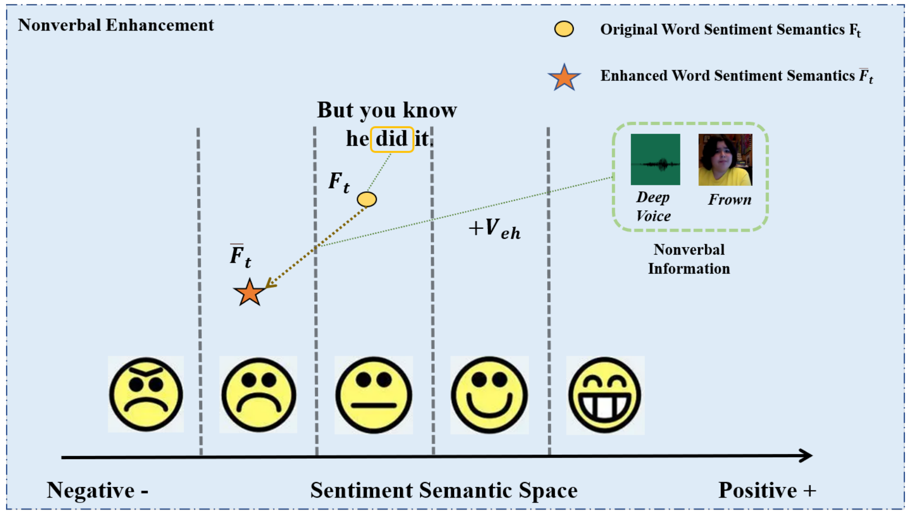


Fig. 7. An example of updating word sentiment semantic representation with nonverbal feature enhancement information. V_{eh} is the nonverbal enhancement embedding. The word "did" in the example is a neutral word, while it can be adjusted to a negative word according to the expression of the relative nonverbal modalities.

$\bar{F}_t$ is passed to the remaining layers of the pre-trained language model. The feature of the textual representation $[CLS]$ from the last layer is embedded into the prediction layer of PLM to predict the sentiment intensity $\hat{y}$. We use L1 Loss as the task loss function, as formula (16) shows,

$$L_{task} = \frac{1}{N_s} \sum_{i=1}^{N_s} |\hat{y}^i - y^i|, \tag{16}$$

where N_s is the number of training samples, and y is the multimodal sentiment label.

4. Experiments

4.1. Experiment Settings

In this section, we detail the experimental setup, including the experimental datasets, the feature extraction, the baseline models, the experimental parameters, and the metrics.

4.1.1. Datasets

We conduct multimodal sentiment analysis experiments on three publicly available datasets, MOSI (Zadeh et al., 2016), MOSEI (Zadeh et al., 2018), and CH-SIMS (Yu et al., 2020). Table 1 presents basic information about the datasets.

MOSI: MOSI (Zadeh et al., 2016) is a multimodal sentiment analysis dataset released in 2016, consisting of 90 YouTube videos and 2199 utterance-level video clips. Each video clip is labelled with sentiment intensity between [-3, 3] and categorized into seven types corresponding to [-3, -2): highly negative, [-2, -1): negative, [-1, 0): weakly negative, [0]: neutral, (0, 1]: weakly positive, (1, 2]: positive, and (2, 3]: highly positive, respectively.

MOSEI: The MOSEI (Zadeh et al., 2018) dataset, released in 2018, is a large-scale multimodal sentiment analysis dataset similar to MOSI. It consists of 3228 videos, which are divided into 22856 short video clips at the utterance level. Each video clip is annotated with multimodal sentiment labels ranging from -3 (strong negative) to 3 (strong positive), which is the same to that of MOSI.

Table 1
Dataset statistics for benchmark MSA datasets in format negative (<0) / neutral (=0) / positive (>0) sentiment intensity.

Dataset	#Train	#Valid	#Test	#Total
MOSI	552/53/679	92/13/124	379/30/277	2199
MOSEI	4738/3540/8084	506/433/932	1350/1025/2284	22856
CH-SIMS	742/207/419	248/69/139	248/69/140	2281

CH-SIMS: CH-SIMS (Yu et al., 2020) is a Chinese multimodal sentiment analysis dataset released in 2020 and contains a total of 60 videos, which are split into 2281 utterance-level short video clips. It has a multimodal sentiment label and three unimodal sentiment labels for each video clip. We only use multimodal sentiment labels in this paper. Unlike CMU-MOSI and CMU-MOSEI datasets, each video clip is labelled with sentiment intensity between [-1, 1] in dataset. The multimodal sentiment labels are categorized into three classifications, where [-1.0, -0.1] represents negative, (-0.1, 0.1] represents neutral, (0.1, 1.0] represents positive. The multimodal sentiment labels are also categorized into five classifications, where [-1.0, -0.7] represents negative, (-0.7, -0.1] represents weakly negative, (-0.1, 0.1] represents neutral, (0.1, 0.7] represents weakly positive, and (0.7, 1.0] represents positive.

4.1.2. Feature Extraction

In order to facilitate a fair comparison with most previous studies, we elaborate on the feature extraction part of different modalities.

For the text sequence, we use the front i-layer of BERT-base to obtain the corresponding intermediate feature sequence F_t. As for the acoustic feature H_a and visual feature H_v, we obtain them by using COVAREP (Degottex et al.,

2014) and OpenFace (Baltrusaitis et al., 2016) for MOSI (Zadeh et al., 2016) and MOSEI (Zadeh et al., 2018), and using LibROSA (McFee et al., 2015) and OpenFace2.0 toolkit (Baltrusaitis et al., 2018) for CH-SIMS (Yu et al., 2020).

4.1.3. Baselines

Numerous methods have been proposed in MSA with Deep Learning (DL), and this paper utilizes several widely adopted baseline models for utterance-level multimodal sentiment analysis (Hazarika et al., 2020; Gandhi et al., 2023), which are tensor based fusion and low-rank variants TFN (Zadeh et al., 2017), LMF (Liu et al., 2018), etc.; graph based fusion models Graph-MFN (Zadeh et al., 2018), etc.; attention and Transformer modules, MulT (Tsai et al., 2019), MAG-BERT (Rahman et al., 2020), HyCon-BERT(Mai et al., 2022), EMT (Sun et al., 2023a), etc.; different interactions of bi-modalities ICCN (Sun et al., 2020), BBFN (Han et al., 2021), etc.; multi-tasking models MISA (Hazarika et al., 2020), Self-MM (Yu et al., 2021), etc.; meta-learning models AMML (Sun et al., 2023b), etc.; models based on MLP-Mixer (Tolstikhin et al., 2021) PS-Mixer (Lin et al., 2023), etc. Description of the baseline models is as follows:

(1) **TFN**: TFN (Zadeh et al., 2017) aggregates unimodal, bimodal, and trimodal interactions using a Cartesian product of tensors.

(2) **LMF**: LMF (Liu et al., 2018) adopts TFN based low-rank factorization to reduce computation and improve model efficiency.

(3) **Graph-MFN**: Graph-MFN (Zadeh et al., 2018) utilizes Dynamic Fusion Graph, which is directly related to how modalities interact.

(4) **MulT**: MulT (Tsai et al., 2019) extends the standard Transformer model to focus on the Cross-Modal interactions of the entire utterance, learning representations from unaligned multimodal data.

(5) **MISA**: MISA (Hazarika et al., 2020) projects each modality into two subspaces to learn modality-invariant and modality-specific representations, and fuses these two representations to predict sentiments.

(6) **MAG-BERT**: MAG-BERT (Rahman et al., 2020) is an improvement over RAVEN (Wang et al., 2019) and incorporates aligned nonverbal information into textual representations by using a multimodal adaptive gating mechanism in a BERT pre-trained model.

(7) **ICCN**: ICCN (Sun et al., 2020) uses pre-trained BERT in a shared semantic space for vision-to-text and audio-to-text translation.

(8) **Self-MM**: Self-MM (Yu et al., 2021) constructs a unimodal label generator, which is based on multi-task learning, self-supervised learning to enhance specific representations of each modality.

(9) **BBFN**: BBFN (Han et al., 2021) strives to properly balance the contributions of different modality pairs using extra loss.

(10) **HyCon-BERT**: (Mai et al., 2022) proposes a novel multimodal representation learning framework HyCon based on contrastive learning, designed with three types of losses to comprehensively learn inter-modal/intra-modal dynamics in both supervised and unsupervised ways.

Table 2
Hyperparameter settings of each dataset.

Hy-Param	MOSI	MOSEI	CH-SIMS
bz	16	64	32
lr	2e-5	2e-5	1e-5
k	16	32	16
a-kz, v-kz	7, 5	7, 3	7, 5
kd	2	2	2
l	1	2	2
a-hs, v-hs	12, 12	12, 12	12, 16
PLM-i	0	1	1

(11) **AMML**: AMML (Sun et al., 2023b) introduces the adaptive multimodal meta-learning mechanism to meta-learn the unimodal networks and adapt them for multimodal inference.

(12) **PS-Mixer**: PS-Mixer (Lin et al., 2023) proposes a mixture model of polarity vector and intensity vector based on the multi-layer perceptron mixture model to achieve better communication between different modality data for multimodal sentiment analysis.

(13) **EMT**: EMT (Sun et al., 2023a) utilizes the efficient multimodal Transformer to better model cross-modal interactions in unaligned multimodal data.

4.1.4. Experimental Parameters and Metrics

Our backbone model is based on the pre-trained BERT-base model. The batch size and learning rate are denoted as "bz" and "lr", respectively. The number of frame-level clustering centers for acoustics and vision is "k". The convolution kernel size, shared expansion coefficient and layers of TCN are denoted as "a-kz", "v-kz", "kd" and "l", respectively. The numbers of heads in the cross-modal attention mechanisms for acoustic and visual modalities are denoted as "a-hs" and "v-hs", respectively. The FNE module is inserted into the backbone model at the "PLM-i" layer. The values of hyperparameters for different datasets are shown in Table 2.

We present the experimental results of our study, which are reported in the form of multi-class classification using methods commonly adopted in the field.

For MOSI and MOSEI, we report 2-class accuracy (Acc-2) and weighted F1 score (F1), and 7-class accuracy (Acc-7). There are two ways to calculate Acc-2 and F1 for MOSI and MOSEI: negative/non-negative (neutral included) (Zadeh et al., 2016) and negative/positive (neutral excluded) (Tsai et al., 2019). As for CH-SIMS, following work (Yu et al., 2020), we report 2-class accuracy (Acc-2) and weighted F1-score (F1), 3-class accuracy (Acc-3), and 7-class accuracy (Acc-7).

For all datasets, Acc-n denotes the ratio that prediction is in the correct interval among the n intervals of the labels. The formula Acc_2 of 2-class accuracy is as follows:

$$Acc_2 = \frac{TP + TN}{N_t}, \tag{17}$$

Table 3
Results of the models on MOSI. *: results from (Sun et al., 2020). #: results from (Rahman et al., 2020). †: results from (Mao et al., 2022). (↑ means higher is better, ↓ means lower is better. In all models, the part in bold is the best result in this column on unaligned and aligned data, respectively. For Acc-2 and F1, the left side of / means "negative/non-negative", and the right is "negative/positive".)

Model	Acc-2(%)↑	F1(%)↑	Acc-7(%)↑	MAE↓	Corr↑	Data setting
TFN*	-/80.82	-/80.77	34.94	0.901	0.698	unaligned
LMF*	-/82.53	-/82.47	33.23	0.917	0.695	unaligned
MulT#	81.50/84.10	80.60/83.90	-	0.861	0.711	unaligned
Self-MM†	-/84.30	-/84.31	-	0.720	0.793	unaligned
PS-Mixer	80.30/82.10	80.30/82.10	44.31	0.794	0.748	unaligned
EMT	83.30/85.00	83.20/85.00	47.40	0.705	0.798	unaligned
FNENet	**83.53/85.52**	**83.45/85.50**	**48.25**	**0.690**	**0.805**	unaligned
Graph-MFN†	-/79.68	-/77.06	-	0.986	0.642	aligned
MISA	81.80/83.40	81.70/83.60	42.30	0.783	0.761	aligned
MAG-BERT†	-/83.41	-/83.47	-	0.761	0.776	aligned
ICCN	-/83.07	-/83.02	39.01	0.862	0.714	aligned
BBFN	-/84.30	-/84.30	45.00	0.776	0.755	aligned
AMML	-/84.90	-/84.80	46.30	0.723	0.792	aligned
HyCon-BERT	-/85.20	-/85.10	46.60	0.713	0.790	aligned
FNENet	**83.53/85.52**	**83.45/85.50**	**48.25**	**0.690**	**0.805**	unaligned

where TP (True Positive) represents the number that the prediction is correctly in the interval of true label in [0, max], TN (True Negative) represents the number that the prediction is correctly in the interval of true label in [min, 0), and N_t represents the total number of test samples.

The $F1$ score is the measure considering both the precision and recall of the model. The formula of weighted $F1$ is as follows:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}, \tag{18}$$

$$Precision = \frac{TP}{TP + FP}, \tag{19}$$

$$Recall = \frac{TP}{TP + FN}, \tag{20}$$

where FP (False Positive) represents the number that the prediction is not in the interval of true label in [0, max], and FN (False Negative) represents the number that the prediction is not in the interval of true label in [min, 0).

We also report the Mean Absolute Error (MAE) and Pearson Correlation (Corr) for the regression task on all datasets. The calculation formula of MAE is the same as L1 Loss. Corr is used to evaluate the linear relationship between two continuous variables, ranging from nonlinear correlation to linear correlation with a value range of [-1, 1]. Except for MAE, higher values of the other metrics indicate better performance. We utilize the AdamW (Loshchilov & Hutter, 2019) optimizer for parameter training and implement the model based on Pytorch framework.

4.2. Result Analysis

Table 3, Table 4, and Table 5 present the experimental results on MOSI, MOSEI, and CH-SIMS datasets, respectively. As MOSI and MOSEI datasets contain aligned and unaligned data, the data settings adopted by each model are marked in the last column of the tables. TFN, LMF, MulT, Self-MM, PS-Mixer, EMT, and FNENet use unaligned data, while Graph-MFN, MISA, MAG-BERT, ICCN, BBFN, AMML, and HyCon-BERT employ aligned data.

Table 4
Results of the models on MOSEI. *: results from (Sun et al., 2020). #: results from (Rahman et al., 2020). †: results from (Mao et al., 2022). (↑ means higher is better, ↓ means lower is better. In all models, the part in bold is the best result in this column on unaligned and aligned data, respectively, and the part with double underline is the sub-optimal result. For Acc-2 and F1, the left side of / means "negative/non-negative", and the right is "negative/positive".)

Model	Acc-2(%)↑	F1(%)↑	Acc-7(%)↑	MAE↓	Corr↑	Data setting
TFN*	- /82.57	-/82.09	50.21	0.593	0.700	unaligned
LMF*	- /82.03	-/82.18	48.02	0.623	0.677	unaligned
MulT#	-/83.50	-/82.90/-	-	-	-	unaligned
Self-MM†	-/84.06	-/84.12	-	$\underline{\underline{0.531}}$	0.766	unaligned
PS-Mixer	83.10/86.10	83.10/86.10	53.00	0.537	0.765	unaligned
EMT	**83.40**/86.00	**83.70**/86.00	**54.50**	**0.527**	**0.774**	unaligned
FNENet	81.05/**86.16**	81.68/**86.20**	$\underline{\underline{54.15}}$	0.536	$\underline{\underline{0.769}}$	unaligned
Graph-MFN†	-/83.48	-/83.23	-	0.575	0.713	aligned
MISA	**83.60**/85.50	83.80/85.30	52.20	0.555	0.756	aligned
MAG-BERT	-/84.70	-/84.50	-	-	-	aligned
ICCN	-/84.18	-/84.15	51.58	0.565	0.713	aligned
BBFN	-/**86.20**	-/86.10	**54.80**	**0.529**	0.767	aligned
AMML	-/85.30	-/85.20	52.40	0.614	0.776	aligned
HyCon-BERT	-/85.40	-/85.60	52.80	0.601	**0.776**	aligned
FNENet	$\underline{\underline{81.05}}$/$\underline{\underline{86.16}}$	$\underline{\underline{81.68}}$/**86.20**	$\underline{\underline{54.15}}$	$\underline{\underline{0.536}}$	$\underline{\underline{0.769}}$	unaligned

Since the text features in some earlier baseline models were extracted with Glove, in order to keep it relatively fair, we use the baseline results of extracting text modality with BERT.

From Table 3, it is obvious that FNENet outperforms all baseline models on all metrics, indicating the superiority of our method on MOSI. Specifically, it surpasses the best-performing method EMT by 0.52% Acc-2, 0.50% F1, 0.015 MAE, and 0.85% Acc-7, and outperforms HyCon-BERT by 0.32% Acc-2, 0.40% F1, 0.023 MAE, and 1.65% Acc-7. From Table 4, we observe that the results of FNENet on MOSEI are either optimal or suboptimal. Specifically, it surpasses the best-performing method EMT by 0.16% Acc-2, 0.20% F1 , and outperforms BBFN by 0.10% F1. We also present the results on the Chinese MSA dataset CH-SIMS in Table 5. It can be seen that FNENet achieves the best performance on most metrics. For instance, it outperforms the second performer EMT by 0.21% Acc-2, 0.28% F1, and 0.92% Acc-7. On these datasets, FNENet's performance is the best compared to the baseline models. We can attribute these encouraging results to the use of vector quantization in FNENet to reduce the inter-modal gap and facilitate fusion. Because all baseline methods focus on the design of effective fusion, they do not take into account the impact of the initial differences between modalities on the fusion stage.

In summary, significant performance improvements to baseline models on three publicly available datasets demonstrate that our FNENet eliminates differences between modalities, promotes the fusion between modalities, and fully uses nonverbal modality information.

4.3. Ablation Study

In this section, we examine the ablation of FNENet on MOSI to investigate the individual contributions of different modules to the overall performance (see Table 6).

The following abbreviations are used:

Table 5
Results of the models on CH-SIMS. *: results from source. †: results from (Mao et al., 2022). (↑ means higher is better, ↓ means lower is better. In all models, the part in bold is the best result in this column, and the part with double underline is the sub-optimal result.)

Model	Acc-2(%)↑	F1(%)↑	Acc-3(%)↑	Acc-5(%)↑	MAE↓	Corr↑
TFN*	78.38	78.62	65.12	39.30	0.432	0.591
LMF*	77.77	77.88	64.68	40.53	0.441	0.576
Graph-MFN*	78.77	78.21	65.65	39.82	0.445	0.578
MulT*	78.56	79.66	64.77	37.94	0.453	0.564
MISA†	76.54	76.59	-	-	0.447	0.563
ICCN	-	-	-	-	-	-
MAG-BERT†	74.44	71.75	-	-	0.492	0.399
Self-MM*	80.04	80.44	65.47	41.53	0.425	0.595
BBFN	-	-	-	-	-	-
AMML	-	-	-	-	-	-
HyCon-BERT	-	-	-	-	-	-
PS-Mixer	-	-	-	-	-	-
EMT	80.10	80.10	67.40	43.50	**0.396**	**0.623**
FNENet	**80.31**	**80.38**	**68.05**	**44.42**	0.417	0.618

Table 6
Performance contributions of different modules of the FNENet model. (↑ means higher is better, ↓ means lower is better. In all models, the part in bold is the best result in this column.)

Model	Acc-2(%)↑	F1(%)↑	MAE↓	Corr↑
FNENet	**85.52**	**85.50**	**0.690**	**0.805**
A_{raw}, V	84.30	84.23	0.707	0.796
A, V_{raw}	84.45	84.44	0.721	0.794
A_{raw}, V_{raw}	82.01	82.10	0.749	0.787
CTC-L	84.30	84.30	0.727	0.793
CTC-G	84.30	84.31	0.718	0.799
w/o CTC-T	82.16	82.18	0.740	0.782
w/o CMA	84.76	84.80	0.716	0.793
w/o SF	82.77	82.84	0.724	0.795
T	84.60	84.63	0.709	0.797
T, A → T	84.15	84.18	0.721	0.794
T, V → T	84.76	84.79	0.722	0.793

- **FNENet**: typical FNENet.
- $\mathbf{A_{raw}, V}$: using raw acoustic modality and keeping visual modality with VQ.
- $\mathbf{A, V_{raw}}$: similar to $\mathbf{A_{raw}, V}$ but exchanging the sources of two modalities.
- $\mathbf{A_{raw}, V_{raw}}$: removal of VQ and both raw acoustic and visual modality are as the input.
- **CTC-L** & **CTC-G**: TCN module in CTC-T replaced by LSTM and GRU, respectively.
- **w/o CTC-T**: removing the TCN based CTC module.
- **w/o CMA**: taking out of the Cross-Modal Attention mechanism.
- **w/o SF**: dislodging the Sequence Fusion module.
- **T**: without integrating nonverbal information, which is equivalent to BERT.
- $\mathbf{T, A \rightarrow T}$: FNENet without integrating visual information.
- $\mathbf{T, V \rightarrow T}$: FNENet without integrating acoustic information.

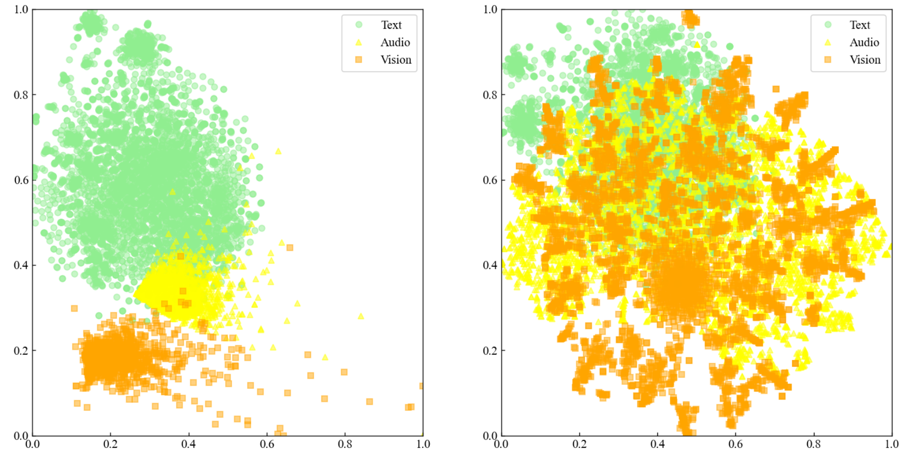
Table 6 shows that removing the VQ of any modality from the FNENet model results in a drastic drop in the model's performance. As previously mentioned, VQ is responsible for converting frame-level features into "word"-level features to reduce the distribution difference of initial features between heterogeneous modalities. Without this module, the distribution difference of heterogeneous data makes subsequent fusion more difficult, severely degrading the model's performance. Additionally, the model's performance declines after removing the CTC-T or replacing TCN with LSTM and GRU. When TCN's ability to model long-distance, local feature dependencies are removed, performance naturally declines. Further, the Cross-Modal Attention is used for the final feature fusion, selectively enhancing nonverbal features with text, and removing it also causes a decline in model performance. Finally, we conduct ablation experiments on the modalities and found that removing any nonverbal modality would result in performance loss, indicating that our model can effectively enhance text modality by utilizing the emotional information of nonverbal modalities. Therefore, this is also the key to improving the accuracy of our model in sentiment analysis.

4.4. Visualization and Analysis

In this section, we conduct a more detailed study on the structure of FNENet. Specifically, we investigate the data visualization (Fig. 8) and present the case study (Table 7).

Fig. 8 shows the data distribution on MOSEI without and with feature transformation strategy by t-SNE (Van der Maaten & Hinton, 2008). In Fig. 8(a), the distribution differences of the three modes in the fusion stage are very obvious. Each of them is clustered at different positions in the figure, and the feature interaction is not obvious. In Fig. 8(b), after using VQ, the distribution of the three modes is very evenly distributed, and there is obvious interaction between them, thus improving the fusion performance.

To demonstrate how FNENet operates, we present examples of sentiment strength prediction with and without nonverbal enhancement embeddings. Table 7 illustrates how FNENet predicts sentiment by incorporating nonverbal information. In the first example, determining sentiment polarity based solely on textual information is inconclusive. In this case, nonverbal information can assist the model in determining the sentiment polarity. In the second and third examples, FNENet predicts sentiment polarity using only textual information, without nonverbal enhancement embeddings, resulting in insufficient predicted sentiment strength. After adding the augmented embedding, the predicted value is much closer to the actual sentiment strength. These observations suggest that FNENet can effectively utilize information from both acoustic and visual modalities to gain a more accurate sentiment prediction.



(a) Without Feature Transformation (Raw Features)

(b) With Feature Transformation (Embedding Features)

Fig. 8. The data distribution of modalities without and with feature transformation by t-SNE.

Table 7
The table shows several examples of sentiment values predicted by unimodal text based on **B**ERT T(B), keeping raw nonverbal feature FNE-R(B), using nonverbal enhancement embedding FNE-E(B) and ground truth (GT). (The bracket (Δ) denotes the absolute value of the difference between the predicted value and the ground truth, with a smaller absolute value indicating better performance. The bold part indicates the best result in each example.)

Modality	Example	T(B)	FNE-R(B)	FNE-E(B)	GT
T	IT ONLY HAD THE POTENTIAL TO BE A FILM THAT WAS BOTH PROVKING AND ACTION PACKED	0.15(Δ0.55)	-0.12(Δ0.28)	**-0.4**(**Δ0.00**)	-0.4
A	Urgent				
V	Normal				
T	NOW THE REAL STAR OF SOMETHING BORROWED IS GINNIFER GOODWIN	0.91(Δ0.29)	0.95(Δ0.25)	**1.29**(**Δ0.09**)	1.20
A	Emphasized				
V	Smile				
T	THATS WHY I WAS NOT EXCITED ABOUT THE FOURTH ONE	-0.95(Δ0.45)	-1.06(Δ0.34)	**-1.33**(**Δ0.07**)	-1.40
A	Low voice				
V	Disappointed				

5. Conclusion

This paper proposes a Frame-level Nonverbal feature Enhancement Network (FNENet) model to enhance textual representations in a pre-trained language model with long-range acoustic and visual sentiment information. A feature transformation mechanism is also introduced to reduce the original distribution difference between verbal and non-verbal modalities. Extensive experiments demonstrate that FNENet outperforms existing baseline models on benchmark datasets MOSI, MOSEI, and CH-SIMS. Future work will explore the impact of different cluster center features on acoustic and visual features, and a more advanced multimodal learning model will be designed to investigate the interaction between verbal and nonverbal features.

Acknowledgement

This study was supported by the Shanghai Service Industry Development Fund and the High Performance Computing Center of Shanghai University, and Shanghai Engineering Research Center of Intelligent Computing System (No. 19DZ2252600).

References

Arandjelovic, R., Gronát, P., Torii, A., Pajdla, T., & Sivic, J. (2018). Netvlad: CNN architecture for weakly supervised place recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, *40*, 1437–1451.

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *CoRR*, *abs/1803.01271*.

Baltrusaitis, T., Robinson, P., & Morency, L. (2016). Openface: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision* (pp. 1–10).

Baltrusaitis, T., Zadeh, A., Lim, Y. C., & Morency, L. (2018). Openface 2.0: Facial behavior analysis toolkit. In *13th IEEE International Conference* *on Automatic Face & Gesture Recognition* (pp. 59–66).

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Cho, K., van Merrienboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1724–1734).

Degottex, G., Kane, J., Drugman, T., Raitio, T., & Scherer, S. (2014). COVAREP - A collaborative voice analysis repository for speech technologies. In *IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 960–964).

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186).

Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., & Hussain, A. (2023). Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Inf. Fusion*, *91*, 424–444.

Graves, A., Fernández, S., Gomez, F. J., & Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006)* (pp. 369–376). volume 148 of *ACM International Conference Proceeding Series*.

Han, W., Chen, H., Gelbukh, A. F., Zadeh, A., Morency, L., & Poria, S. (2021). Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In *ICMI '21: International Conference on Multimodal Interaction* (pp. 6–15).

Hausler, S., Garg, S., Xu, M., Milford, M., & Fischer, T. (2021). Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 14141–14152).

Hazarika, D., Zimmermann, R., & Poria, S. (2020). MISA: modality-invariant and -specific representations for multimodal sentiment analysis. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle* (pp. 1122–1131).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, *9*, 1735–1780.

Jégou, H., Douze, M., Schmid, C., & Pérez, P. (2010). Aggregating local descriptors into a compact image representation. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3304–3311).

Lin, H., Zhang, P., Ling, J., Yang, Z., Lee, L. K., & Liu, W. (2023). Ps-mixer: A polar-vector and strength-vector mixer model for multimodal sentiment analysis. *Inf. Process. Manag.*, *60*, 103229.

Lin, R., Xiao, J., & Fan, J. (2018). Nextvlad: An efficient neural network to aggregate frame-level features for large-scale video classification. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops* (pp. 0–0).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, *abs/1907.11692*.

Liu, Z., Shen, Y., Lakshminarasimhan, V. B., Liang, P. P., Zadeh, A., & Morency, L. (2018). Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers* (pp. 2247–2256).

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3431–3440).

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*.

Mai, S., Zeng, Y., Zheng, S., & Hu, H. (2022). Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, .

Mao, H., Yuan, Z., Xu, H., Yu, W., Liu, Y., & Gao, K. (2022). M-sena: An integrated platform for multimodal sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 204–213).

McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science Conference 2015 (SciPy 2015)* (pp. 18–24).

Peng, J., Wu, T., Zhang, W., Cheng, F., Tan, S., Yi, F., & Huang, Y. (2023). A fine-grained modal label-based multi-stage network for multimodal sentiment analysis. *Expert Systems with Applications*, *221*, 119721.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1532–1543).

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 2227–2237).

Rahman, W., Hasan, M. K., Lee, S., Zadeh, A. B., Mao, C., Morency, L., & Hoque, M. E. (2020). Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 2359–2369).

Salimans, T., & Kingma, D. P. (2016). Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016* (p. 901).

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, *15*, 1929–1958.

Sun, L., Lian, Z., Liu, B., & Tao, J. (2023a). Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, .

Sun, Y., Mai, S., & Hu, H. (2023b). Learning to learn better unimodal representations via adaptive multimodal meta-learning. *IEEE Transactions on Affective Computing*, *14*, 2209–2223.

Sun, Z., Sarma, P. K., Sethares, W. A., & Liang, Y. (2020). Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence* (pp. 8992–8999).

Tolstikhin, I. O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., Lucic, M., & Dosovitskiy, A. (2021). Mlp-mixer: An all-mlp architecture for vision. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021* (pp. 24261–24272).

Tsai, Y. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L., & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, Volume 1: Long Papers* (pp. 6558–6569).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *CoRR*, *abs/1706.03762*.

Wang, D., Liu, S., Wang, Q., Tian, Y., He, L., & Gao, X. (2022). Cross-modal enhancement network for multimodal sentiment analysis. *IEEE Transactions on Multimedia*, (pp. 1–13).

Wang, X., Zhu, L., & Yang, Y. (2021). T2VLAD: global-local sequence alignment for text-video retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5079–5088).

Wang, Y., Shen, Y., Liu, Z., Liang, P. P., Zadeh, A., & Morency, L. (2019). Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *The Thirty-Third AAAI Conference on Artificial Intelligence* (pp. 7216–7223).

Wang, Z., Wan, Z., & Wan, X. (2020). Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis. *CoRR*, *abs/2009.02902*.

Wu, T., Peng, J., Zhang, W., Zhang, H., Tan, S., Yi, F., Ma, C., & Huang, Y. (2022). Video sentiment analysis with bimodal information-augmented multi-head attention. *Knowl. Based Syst.*, *235*, 107676.

Xu, Q., Peng, J., Zheng, C., Tan, S., Yi, F., & Cheng, F. (2023). Short text classification of chinese with label information assisting. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *22*, 1–19.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019* (pp. 5754–5764).

Yu, F., & Koltun, V. (2016). Multi-scale context aggregation by dilated convolutions. In *4th International Conference on Learning Representations*.

Yu, W., Xu, H., Meng, F., Zhu, Y., Ma, Y., Wu, J., Zou, J., & Yang, K. (2020). CH-SIMS: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 3718–3727).

Yu, W., Xu, H., Yuan, Z., & Wu, J. (2021). Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Thirty-Fifth AAAI Conference on Artificial Intelligence* (pp. 10790–10797).

Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. (2017). Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 1103–1114).

Zadeh, A., Liang, P. P., Poria, S., Cambria, E., & Morency, L. (2018). 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* (pp. 2236–2246).

Zadeh, A., Mao, C., Shi, K., Zhang, Y., Liang, P. P., Poria, S., & Morency, L. (2019). Factorized multimodal transformer for multimodal sequential learning. *CoRR*, *abs/1911.09826*.

Zadeh, A., Zellers, R., Pincus, E., & Morency, L. (2016). MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *CoRR*, *abs/1606.06259*.

Zhao, T., Peng, J., Huang, Y., Wang, L., Zhang, H., & Cai, Z. (2023). A graph convolution-based heterogeneous fusion network for multimodal sentiment analysis. *Applied Intelligence*, *53*, 30455–30468.