DOI: 10.19884/j.1672-5220.202402011

A Dual Stream Multimodal Alignment and Fusion Network for Classifying Short Videos

ZHOU Ming[1,2], WANG Tong[1,2*]
1. College of Information Science and Technology, Donghua University, Shanghai 201620, China
2. Engineering Research Center of Digitized Textile & Apparel Technology, Ministry of Education, Donghua University, Shanghai 201620, China

Abstract: Video classification is an important task in video understanding and plays a pivotal role in intelligent monitoring of information content. Most existing methods do not consider the multimodal nature of the video, and the modality fusion approach tends to be too simple, often neglecting modality alignment before fusion. This research introduces a novel dual stream multimodal alignment and fusion network named DMAFNet for classifying short videos. The network uses two unimodal encoder modules to extract features within modalities and exploits a multimodal encoder module to learn interaction between modalities. To solve the modality alignment problem, contrastive learning is introduced between two unimodal encoder modules. Additionally, masked language modeling (MLM) and video text matching (VTM) auxiliary tasks are introduced to improve the interaction between video frames and text modalities through backpropagation of loss functions. Diverse experiments prove the efficiency of DMAFNet in multimodal video classification tasks. Compared with other two mainstream baselines, DMAFNet achieves the best results on the 2022 WeChat Big Data Challenge dataset.

Keywords: video classification; multimodal fusion; feature alignment
CLC number: TP751.1 **Document code**: A
Article ID: 1672-5220(2025)01-0088-08

Open Science Identity (OSID)

0 Introduction

In recent years, the popularity of smartphones and the widespread use of social media have led to the sharing of a large amount of user-generated content (UGC), most of which is in the form of short videos. Doing a good job of classifying short videos also helps with the intelligent monitoring of online content. As a result, short video classification has become an important task in the video field. However, it still faces challenges in achieving accurate classification results in short videos that contain multiple modalities. One challenge is modality alignment. Since different modality features may have different sampling rates, different modality features need to be mapped to the same frequency. The other is modality fusion. The fused features play a decisive role in the classification results.

Traditional video classification methods primarily describe videos by extracting features of key points, which are mainly divided into spatiotemporal key points[1] and dense trajectories[2]. The core concept based on spatiotemporal key points is that key points in video frames are usually data that change drastically in the spatiotemporal dimension, and these data often reflect important information about target movements. The success of video classification relies on the information of these target movements. Another classic method is the dense trajectories method, which tracks the changes of a given coordinate image over time to capture the target's motion information. However, these methods require manual feature extraction which not only requires a lot of extra labor but also results in low accuracy. Recently, researchers have made significant progress in video classification methods, largely due to the development of deep learning, such as convolutional neural networks (CNNs). The current video classification methods based on deep learning mainly include recurrent neural networks (RNNs)/long short-term memory (LSTM) sequential models[3], three-dimensional (3D) convolutional networks[4] and attention mechanism[5].

To learn the temporal information in videos, researchers have expanded many mainstream image classification methods to the temporal dimension and proposed many 3D convolutional networks. The conventional 3D (C3D) network[6-7] aligns with VGG16, incorporating eight convolutional layers with a convolution kernel size of $3 \times 3 \times 3$ and five pooling

Received date: 2024-02-27
Foundation items: Fundamental Research Funds for the Central Universities, China (No. 2232021A-10); National Natural Science Foundation of China (No. 61903078); Shanghai Sailing Program, China (No. 22YF1401300); Natural Science Foundation of Shanghai, China (No. 20ZR1400400)
* Correspondence should be addressed to WANG Tong, email: wangtong@dhu.edu.cn
Citation: ZHOU M, WANG T. A dual stream multimodal alignment and fusion network for classifying short videos [J]. *Journal of Donghua University (English Edition)*, 2025, 42(1): 88-95.

layers. Based on InceptionV1, Carreira et al.[8] extended the original convolution and pooling kernel from two dimensions to three dimensions, initialized the network with the original parameters, and finally achieved a very deep spatiotemporal classification network. Because of the additional time dimension, the number of parameters and calculations of 3D convolutional networks increase exponentially compared with two-dimensional (2D) convolutional networks. Therefore, the separable 3D (S3D) convolutional network proposed by DeepMind[9] and the R(2+1)D model proposed by Facebook Research[10] both adopt the convolution kernel decomposition idea, decomposing 3D convolutions into time and space dimensions. The former is based on Inception, and the latter is based on ResNet.

With the success of transformers in the fields of natural language processing and vision, some researchers also introduced transformers into the field of video understanding. Arnab et al.[11] proposed a video vision transformer that was composed of two encoders in time and space dimensions. TimeSformer[12] divided attention into space-time attention. Li et al.[13] proposed the UniTransformer model by combining 3D convolutional networks and attention mechanisms. Dave et al.[14] divided video frame features into time-invariant and time-distinctive parts, and used distillation to improve classification results. He et al.[15] combined video tasks with large language models and proposed a long video understanding algorithm, which achieved good results on multiple video tasks.

In this research, we propose a novel dual stream multimodal alignment and fusion network (DMAFNet) for classifying multimodal videos. Compared with previous methods, we take text modality features into account. First, in the extraction of modality features, unimodal encoders and multimodal encoders are used to learn the relationships within modalities and between modalities, respectively. Then, the network achieves spatial alignment of modality features by adding contrastive learning between unimodal encoders. Finally, masked language modeling (MLM) and video text matching (VTM) auxiliary tasks are used in the network to gain more fine-grained interaction between video frames and text modalities through backpropagation. To validate the performance of our proposed network, we conduct comprehensive experiments on the 2022 WeChat Big Data Challenge dataset. The major contributions of this research are as follows.

1) An efficient video classification network DMAFNet is proposed. This network not only models the relationships within modalities but also learns the interactions between modalities. With the help of the cross-attention (CA) mechanism, the network combines text features to obtain video frame features that contain semantic information.

2) A contrastive learning is designed between two unimodal encoders. This method spatially aligns the video frame modality and text modality to facilitate subsequent fusion operations.

3) MLM and VTM auxiliary tasks are introduced to improve the interaction between video frame modality and text modality through backpropagation.

We organize the rest of the paper as follows. Section 1 introduces the architecture and specific components of DMAFNet. Section 2 shows the results and analyses of the comparison experiment and the ablation experiment. Section 3 presents the conclusions.

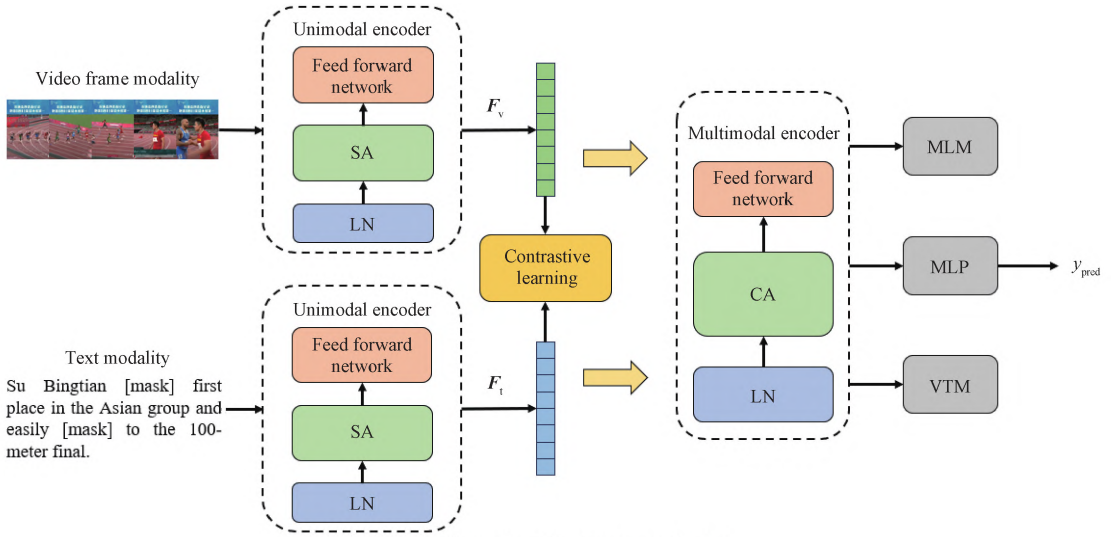
1 Methodology

The architecture of DMAFNet is shown in Fig. 1, which contains three components.

1) Unimodal encoder module. This module is used to extract the features of each modality and learn the relationships within each modality, focusing on modeling within the modality. We exploit the pre-trained bidirectional encoder representations from transformers (BERT) model to process text modality. As for the extraction of video frame modality features, we use the transformer. With the help of these two unimodal encoders, we can obtain the video frame features $\boldsymbol{F}_{\mathrm{v}}$ and text features $\boldsymbol{F}_{\mathrm{t}}$.

2) Multimodal encoder module. This module is used to learn the relationship between text modality and video frame modality, focusing on modeling between modalities. We use a CA mechanism to handle the fusion of modalities. Specifically, we use text modality features as query and video frame modality features as key and value, respectively.

3) Auxiliary tasks and contrastive learning. With the help of contrastive learning, we can align the two modalities before modality fusion, which helps to obtain better fusion features. Two auxiliary tasks, MLM and VTM, are also designed to better model the relationship between modalities. After obtaining the final fused features, we input them into a multi-layer perceptron (MLP) to predict the category of the short video. We use y_{pred} to represent the final prediction result.



SA—self-attention; LN—layer normalization.

Fig. 1 Architecture of DMAFNet

1.1 Problem statement

Multimodal short video classification is one of the basic technologies in the field of video understanding. It has a very wide range of applications in security auditing, recommendation operation, content searching and other fields. A short video usually contains three types of modality information, namely text, audio and video, which play an important role in the mutual promotion and complementation of classification systems at different semantic levels. In this work, multimodal short video classification uses raw text and visual information as inputs, representing text modality and video frame modality, respectively. The goal of our network is to thoroughly model and learn intra-modal and inter-modal relationships, and to achieve good alignment between two modalities. The ultimate objective is to generate a multimodal fusion feature that can be used to accurately predict the category of short videos.

1.2 Data preprocessing

1) Video frame modality. Considering that it is a short video, the video duration in the dataset is generally less than 1 min. Therefore, we sample at a sampling rate of one frame per second and only retain the first 32 frames. If the short video is less than 32 s, all sampled images will be retained. In the end, we get up to 32 images from each short video. Then, we process the obtained images through the vision transformer (ViT) model which is pretrained on Kinetics-400 datasets to obtain the corresponding video frame modality features $\boldsymbol{X}_{\mathrm{v}} \in R^{L\times d}$, where L means sequence length, and $L=32$; d denotes the output dimension of the pre-trained model.

2) Text modality. The text modality in short videos generally consists of title, optical character recognition (OCR) and automatic speech recognition (ASR). Reasonable utilization of the text modality has a great impact on the final classification results. First, we clean the text data and remove special characters and some punctuation marks. Then, we concatenate these texts according to their importance. Usually, the title of a video can reflect the content of the video, so we put the title first, and concatenate ASR and OCR in order according to the importance of the content. Finally, considering that the longer the text length, the greater the calculation amount, we intercept the first 256 tokens as the input of the subsequent text encoder.

1.3 Unimodal encoder module

To gain enhanced feature representations for the video frame modality, we initially perform an LN on the feature $\boldsymbol{X}_{\mathrm{v}} \in R^{L\times d}$: $\boldsymbol{F}_{\mathrm{v}}^{[i]}=\mathrm{LN}(\boldsymbol{F}_{\mathrm{v}}^{[i]})$, where $\boldsymbol{F}_{\mathrm{v}}^{[0]}=\boldsymbol{X}_{\mathrm{v}}$. Following this, we put $\boldsymbol{F}_{\mathrm{v}}^{[i]}$ into an SA block,

$$\boldsymbol{F}_{\mathrm{v}}^{[i+1]}=\mathrm{SA}_{\mathrm{v}}^{[i]}(\boldsymbol{F}_{\mathrm{v}}^{[i]})=\mathrm{softmax}\left(\frac{\boldsymbol{F}_{\mathrm{v}}^{[i]}\boldsymbol{W}_{Q_{\mathrm{v}}}\boldsymbol{W}_{K_{\mathrm{v}}}^{\mathrm{T}}\boldsymbol{F}_{\mathrm{v}}^{[i]\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{F}_{\mathrm{v}}^{[i]}\boldsymbol{W}_{V_{\mathrm{v}}},\tag{1}$$

where $\boldsymbol{W}_{*}$ means learnable weight matrices related to query, key and value of video frame modality; the superscript $[i]$ indicates the number of layers.

After that, we process $\boldsymbol{F}_{\mathrm{v}}^{[i+1]}$ by a position-wise feed forward network (FFN) with residual connections:

$$\boldsymbol{F}_{\mathrm{v}}^{[i+1]}=\mathrm{FFN}(\mathrm{LN}(\boldsymbol{F}_{\mathrm{v}}^{[i+1]}))+\boldsymbol{F}_{\mathrm{v}}^{[i+1]}.\tag{2}$$

Like the transformer, we stack a total of 12 video frame encoders.

Finally, we take the output of the last layer as the video frame features $\boldsymbol{F}_{\mathrm{v}}$. As for the text modality, we follow the previous methods to utilize the pre-trained BERT[16] to extract sentence-level text features $\boldsymbol{F}_{\mathrm{t}}$,

$$\boldsymbol{F}_{\mathrm{t}}=\mathrm{BERT}(\boldsymbol{X}_{\mathrm{t}};\theta_{\mathrm{t}}^{\mathrm{bert}}),\tag{3}$$

where $\boldsymbol{X}_{\mathrm{t}}$ means initial text features; $\theta_{\mathrm{t}}^{\mathrm{bert}}$ denotes the parameters of BERT.

1.4 Multimodal encoder module

Data from different modalities often have unique

information that can complement each other to improve video classification performance. Video frame modality usually includes visual features, such as objects, scenes and colors. Text modality provides semantic information about objects. Video frame modality combined with text modality through a CA mechanism provides video features containing semantic information.

The architecture of the multimodal encoder is shown in Fig. 2. In our proposed network, the multimodal encoder module is stacked N times. In the following ablation experiments, we test the impact of the number of stacks on the final classification results. The module contains LN, a CA block and FFN. The CA block takes $\boldsymbol{F}_{\mathrm{t}}$ as query and takes $\boldsymbol{F}_{\mathrm{v}}$ as key and value. $\boldsymbol{V}$, $\boldsymbol{K}$ and $\boldsymbol{Q}$ stand for three vectors of value, key and query, respectively; $\boldsymbol{F}_{\mathrm{vt}}$ represents the fused features.

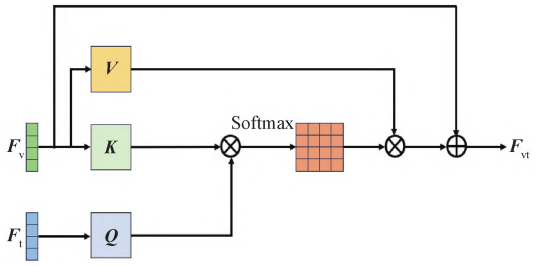


Fig. 2 Architecture of multimodal encoder

First, we perform LN on the features: $\boldsymbol{F}_{\mathrm{v}} = \mathrm{LN}(\boldsymbol{F}_{\mathrm{v}})$ and $\boldsymbol{F}_{\mathrm{t}} = \mathrm{LN}(\boldsymbol{F}_{\mathrm{t}})$. Then, we put them into the CA block,

$$\begin{aligned}\boldsymbol{F}_{\mathrm{vt}}^{[i+1]} &= \mathrm{CA}_{\mathrm{vt}}^{[i]}(\boldsymbol{F}_{\mathrm{vt}}^{[i]}, \boldsymbol{F}_{\mathrm{t}}) \\ &= \mathrm{softmax}\left(\frac{\boldsymbol{F}_{\mathrm{vt}}^{[i]}\boldsymbol{W}_{Q_{\mathrm{t}}}\boldsymbol{W}_{K_{\mathrm{v}}}^{\mathrm{T}}\boldsymbol{F}_{\mathrm{vt}}^{[i]\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{F}_{\mathrm{vt}}^{[i]}\boldsymbol{W}_{V_{\mathrm{v}}}, \end{aligned} \tag{4}$$

where $\boldsymbol{F}_{\mathrm{vt}}^{[0]} = \boldsymbol{F}_{\mathrm{v}}$. After that, we process $\boldsymbol{F}_{\mathrm{vt}}^{[i+1]}$ by a position-wise FFN with residual connections,

$$\boldsymbol{F}_{\mathrm{vt}}^{[i+1]} = \mathrm{FFN}(\mathrm{LN}(\boldsymbol{F}_{\mathrm{vt}}^{[i+1]})) + \boldsymbol{F}_{\mathrm{vt}}^{[i]}. \tag{5}$$

With the help of the CA mechanism, the fused features are constrained to the semantic information which is important for video understanding, thereby reducing the redundant features of the video frame modality. Subsequently, $\boldsymbol{F}_{\mathrm{vt}}$ is fed into an MLP to predict y_{pred} of short videos,

$$y_{\mathrm{pred}} = \mathrm{MLP}(\boldsymbol{F}_{\mathrm{vt}}, \theta_{\mathrm{mlp}}), \tag{6}$$

where θ_{mlp} denotes the parameters of MLP.

1.5 Contrastive learning and auxiliary task

As shown in Fig. 1, we add a contrastive learning module between two unimodal encoder modules. First, the contrastive learning module brings positively correlated video-text pairs closer together in the feature space and obtains better unimodal features before fusion. Then, we use the classification head from the output of the last layer of the unimodal encoder as the input of the contrastive learning module to calculate the contrastive loss for the video-text pairs. Finally, this module aligns video frame and text modalities in the feature space by minimizing the loss function.

In addition to contrastive loss for video-text alignment, we design MLM and VTM auxiliary tasks to motivate fine-grained interaction between video frame and text modalities. The MLM auxiliary task predicts the masked words based on the output of the multimodal encoder. Specifically, we randomly mask 15% of the text words. Out of those masked words, 80% of the words are replaced with [mask], 10% of the words are replaced randomly, and the remaining 10% of the words remain unchanged. Unlike the BERT model, the prediction task not only uses the semantic information of the text but also combines video frame features. By leveraging video frame features to predict masked words, the connection between video frame features and text features is strengthened.

The VTM auxiliary task determines whether the pair of video frame features $\boldsymbol{F}_{\mathrm{v}}$ and text features $\boldsymbol{F}_{\mathrm{t}}$ is matched. This pair is then fed into a multimodal encoder and the first output word embedding is regarded as the joint video-text representation V_i and passes through a fully connected layer for binary prediction. The VTM auxiliary task allows the model to learn whether the input video-text pairs match, and further narrow the relationship between the videos and texts through the backpropagation of the loss function.

1.6 Loss function

The loss function consists of four parts: the classification loss l_{class}, the contrastive loss l_{con}, the MLM loss l_{mlm} and the VTM loss l_{vtm}. The composition of the overall loss function L is

$$L = l_{\mathrm{class}} + l_{\mathrm{con}} + \alpha l_{\mathrm{mlm}} + \beta l_{\mathrm{vtm}}, \tag{7}$$

where α and β are variable hyperparameters, and we set $\alpha = 0.25$ and $\beta = 0.25$ in this paper.

l_{class} is calculated by using the cross-entropy loss between the predicted values of the network output and ground truth values:

$$l_{\mathrm{class}} = \mathrm{CE}(y_{\mathrm{pred}}, y_{\mathrm{gt}}), \tag{8}$$

where CE means cross-entropy loss function; y_{gt} refers to the ground truth values.

We use multiple instance learning noise contrastive estimation (MIL-NCE) to calculate l_{con} of the output of two unimodal encoders:

$$l_{\mathrm{con}} = -\log\left(\frac{\sum_{\boldsymbol{F}_{\mathrm{t}}, \boldsymbol{F}_{\mathrm{v}} \in p}\exp(\boldsymbol{F}_{\mathrm{t}}^{\mathrm{T}}\boldsymbol{F}_{\mathrm{v}}/\tau)}{\sum_{\boldsymbol{F}_{\mathrm{t}}, \boldsymbol{F}_{\mathrm{v}} \in p}\exp(\boldsymbol{F}_{\mathrm{t}}^{\mathrm{T}}\boldsymbol{F}_{\mathrm{v}}/\tau) + \sum_{\boldsymbol{F}'_{\mathrm{t}}, \boldsymbol{F}'_{\mathrm{v}} \in n}\exp(\boldsymbol{F}'^{\mathrm{T}}_{\mathrm{t}}\boldsymbol{F}'_{\mathrm{v}}/\tau)}\right), \tag{9}$$

where p means positive video-text pairs; n means negative video-text pairs; τ is a learnable temperature parameter;

$\boldsymbol{F}'_{\mathrm{t}}$ and $\boldsymbol{F}'_{\mathrm{v}}$ represent the text features and video frame features in the negative video-text pairs, respectively. With the help of this loss function, we can get the aligned features before feature fusion.

The MLM auxiliary task utilizes both video frame modality and text modality to predict the masked words. Supposing that T_i refers to masked words and $p_{\mathrm{mlm}}(V_i, T_i)$ means the prediction result of the masked words, the loss function of MLM can be expressed as

$$l_{\mathrm{mlm}} = -\frac{1}{K}\sum_{i}^{K}\sum_{v}^{V} y_{\mathrm{mlm}}\log_2(p_{\mathrm{mlm}}(V_i, T_i)), \quad (10)$$

where y_{mlm} is a sign function which has the value of 1 if the masked words can be predicted correctly; V is the vocabulary size; K is the number of video-text pairs in the MLM auxiliary task.

The VTM auxiliary task predicts whether video-text pairs match. We define the prediction as $p_{\mathrm{vtm}}(V_i, T_i)$, the loss function of VTM can be expressed as

$$l_{\mathrm{vtm}} = -\frac{1}{M}\sum_{i}^{M}\sum_{t=0}^{1} y_{\mathrm{vtm}}\log_2(p_{\mathrm{vtm}}(V_i, T_i)), \quad (11)$$

where y_{vtm} is a sign function which has the value of 1 if video-text pairs match else 0 if video-text pairs do not match; M is the number of video-text pairs in the VTM auxiliary task. The two auxiliary tasks of VTM and MLM are incorporated to motivate fine-grained interaction between video and text.

2 Experiments and Analyses

2.1 Dataset

The 2022 WeChat Big Data Challenge dataset which adopts a specific data format is shown in Table 1. This dataset collects short video data from WeChat video accounts and contains 100 000 pieces of annotated data. We divide the dataset into 90 000 training sets and 10 000 test sets at a ratio of 9∶1. For specific operations on the dataset, please see Section 1.2.

Table 1 Composition of the dataset

Composition	Explanation
id	Unique identifier of the video
category_id	Manually labeled video classification identity
title	Short video title
frame	First 32 frames of video
asr	Audio to text in video
ocr	Optical character recognition in video

2.2 Evaluation metrics

To assess the accuracy of the network's prediction, we use the F1 score. Since there are multiple categories and the categories are unbalanced, we use both F1 micro $F_{1,\mathrm{mic}}$ and F1 macro $F_{1,\mathrm{mac}}$ and take the average. At the same time, the classification system includes first-level classification and second-level classification, which are calculated separately and averaged during the evaluation. The evaluation metrics are

$$P_{\mathrm{mic}} = \frac{T_{\mathrm{TP}}}{T_{\mathrm{TP}} + T_{\mathrm{FP}}}, \quad (12)$$

$$R_{\mathrm{mic}} = \frac{T_{\mathrm{TP}}}{T_{\mathrm{TP}} + T_{\mathrm{FN}}}, \quad (13)$$

$$F_{1,\mathrm{mic}} = \frac{2(P_{\mathrm{mic}} \times R_{\mathrm{mic}})}{P_{\mathrm{mic}} + R_{\mathrm{mic}}}, \quad (14)$$

$$P_{\mathrm{mac}} = \frac{\sum_{i=1}^{n} P_i}{n}, \quad (15)$$

$$R_{\mathrm{mac}} = \frac{\sum_{i=1}^{n} R_i}{n}, \quad (16)$$

$$F_{1,\mathrm{mac}} = \frac{2(P_{\mathrm{mac}} \times R_{\mathrm{mac}})}{P_{\mathrm{mac}} + R_{\mathrm{mac}}}, \quad (17)$$

where P_{mic} represents the micro precision; R_{mic} represents the micro recall; P_{mac} represents the macro precision; R_{mac} represents the macro recall; T_{TP} represents the number of true positive samples; T_{FP} stands for the number of false positive samples; T_{TN} denotes the number of true negative samples; T_{FN} signifies the number of false negative samples. We obtain four F1 scores: category1_f_1_micro $f_{1,\mathrm{mic}}$, category1_f_1_macro $f_{1,\mathrm{mac}}$, category2_f_2_micro $f_{2,\mathrm{mic}}$ and category2_f_2_macro $f_{2,\mathrm{mac}}$. These four F1 scores represent the $F_{1,\mathrm{mic}}$ and $F_{1,\mathrm{mac}}$ of category 1 and the $F_{1,\mathrm{mic}}$ and $F_{1,\mathrm{mac}}$ of category 2, respectively. The final score F_{sco} is the average of the above four F1 scores, which is shown as

$$F_{\mathrm{sco}} = \frac{1}{4}(f_{1,\mathrm{mic}} + f_{1,\mathrm{mac}} + f_{2,\mathrm{mic}} + f_{2,\mathrm{mac}}). \quad (18)$$

In all the following experiments, we use F_{sco} to represent the accuracy of short video classification.

2.3 Implementation details

During the training process, we use the PyTorch framework to implement the method and train our network on two NVIDIA 3090TI GPUs with 24 GB memory. The network is trained using the Adam optimizer with 50 epochs. The initial learning rate is set to 5×10^{-5}, and the weight decay rate is set to 0.01. Additionally, during training, we set the batch size to 80. The length of the video frame sequence is set to 32 and the length of the text sequence is set to 128.

2.4 Experimental results

We validate the network by conducting comprehensive experiments. In these experiments, our method is compared with the current mainstream short video classification methods, NeXtVALD and VideoBERT, in the industry.

1) Baseline. Currently, the classification of short video in industry mainly uses the NeXtVALD module to

process the fusion of video frame features, and then splices them with the text features output by BERT. The spliced features are sent to a linear layer to obtain the classification result.

2) Quantitative comparison. Table 2 shows the results of different short video classification methods on the 2022 WeChat Big Data Challenge dataset. The arrow in Table 2 indicates that the higher the value of F_{sco}, the better the performance of methods. The best results are in bold. Compared with other methods, DMAFNet does feature alignment work before feature fusion. Moreover, DMAFNet not only learns the relationships between modalities but also the relationships within modalities.

The results of NeXtVALD[17] and VideoBERT[18] are trained using the 2022 WeChat Big Data Challenge dataset. DMAFNet-pretrain means that we use additional unlabeled data to pretrain models on MLM and VTM auxiliary tasks and finetune the 2022 WeChat Big Data Challenge dataset.

Table 2 F_{sco} on 2022 WeChat Big Data Challenge dataset

Method	F_{sco} ↑
NeXtVALD[17]	0.578 2
VideoBERT[18]	0.627 1
DMAFNet (ours)	**0.657 3**
DMAFNet-pretrain (ours)	**0.682 4**

Compared with other methods, DMAFNet takes the multimodal nature of videos into account. Based on the semantic information of text modality, DMAFNet understands the content of the video better and improves the accuracy of video classification. With the help of MLM and VTM auxiliary tasks, DMAFNet gains more fine-grained fused features, and therefore it improves results in multi-level classification datasets like the 2022 WeChat Big Data Challenge dataset. As illustrated in Table 2, it is obvious that DMAFNet achieves the best result compared with the other two mainstream methods in fine-grained classification of multimodal short video.

2.5 Ablation studies

The overall outstanding performance of DMAFNet has confirmed its superiority. To further demonstrate the importance of its various parts, we conduct ablation experiments on the 2022 WeChat Big Data Challenge dataset. The results of these experiments are shown in Tables 3–5.

We separately remove each modality to investigate the significance of multimodality. In Table 3, V means only using video frame modality; T means only using text modality; V+T means using the joint modality of video frame and text. As illustrated in Table 3, it can be seen that the result of V+T has a significant improvement compared to the result of only using V or T modality. With the help of the CA mechanism, the final joint features focus more on information that is beneficial to video classification, such as people, objects and action information.

Table 3 Ablation study on the influence of modality

Input	F_{sco} ↑
T	0.508 2
V	0.551 6
V+T	**0.657 3**

The importance of the proposed components is explored by removing l_{con}, l_{mlm} and l_{vtm} separately. In Table 4, √ refers to the addition of the corresponding loss function. It can be seen that the score degrades when removing the loss functions. These observations suggest that these loss functions play an important role in the results. Specifically, under the constraints of l_{con}, the two modalities can be spatially aligned. The functions of l_{mlm} and l_{vtm} are to strengthen the interaction between the two modalities.

Table 4 Ablation study on the influence of loss function

Method	l_{con}	l_{mlm}	l_{vtm}	F_{sco} ↑
DMAFNet	√	√	√	**0.657 3**
DMAFNet	√	√		0.651 9
DMAFNet	√			0.646 8
DMAFNet				0.635 1

To further verify the effectiveness of DMAFNet, the robustness experiments are conducted on the 2022 WeChat Big Data Challenge dataset. We randomly drop 30% of video frame modality features and text modality features to simulate the absence of some modality features. We also add Gaussian noise to the video frame modality features to verify the robustness of the DMAFNet. As shown in Table 5, it can be seen that although DMAFNet has experienced a decline in classification results in these two cases, the degree of the decline is much smaller compared to the decline observed in the other two mainstream methods. It can be concluded that DMAFNet is relatively robust when modality features are partially lost or affected by additional noise.

Table 5 Results of robustness experiment

Experiment	Method	F_{sco} ↑
Drop 30% features	NeXtVALD[17]	0.519 4
	VideoBERT[18]	0.577 6
	DMAFNet (ours)	**0.618 5**
	DMAFNet-pretrain (ours)	**0.650 4**
Add noise	NeXtVALD[17]	0.498 8
	VideoBERT[18]	0.576 5
	DMAFNet (ours)	**0.617 6**
	DMAFNet-pretrain (ours)	**0.658 1**

To investigate the impact of the number of the multimodal encoders N, experiments on the 2022 WeChat Big Data Challenge dataset are carried out using diverse values for parameter N. The results are illustrated in Fig. 3. As N increases, a curve in which the F_{sco} initially increases and then decreases is observed, with DMAFNet achieving its optimal performance when $N = 6$. The experiments indicate that an excessive number of multimodal encoders may lead to overfitting of network parameters. We should choose the appropriate number of multimodal encoders in our network for different datasets.

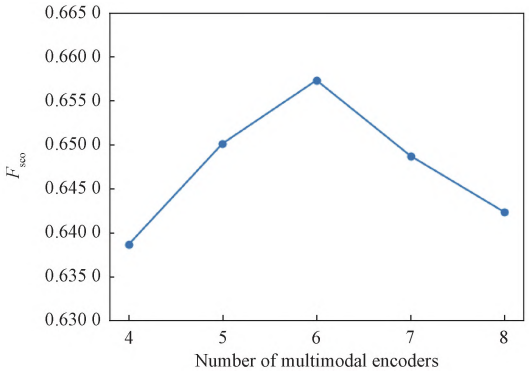


Fig. 3 Performance of DMAFNet with different number of multimodal encoders

3 Conclusions

In this research, we propose DMAFNet for classifying short videos. It mainly consists of two unimodal encoder modules and a multimodal encoder module. The role of the unimodal encoder is to model and learn the relationship within each modality. Moreover, the multimodal encoder is used to perform feature fusion and model the relationship between the modalities. Besides, we add multiple loss functions to DMAFNet to obtain a fine-grained fusion feature. The contrastive loss is to align features before feature fusion. In addition to the contrastive loss for video frame features and text features alignment, VTM loss and MLM loss are incorporated to motivate fine-grained interaction between video and text. DMAFNet rationally utilizes multiple modalities and effectively solves the challenges related to modal alignment and modal fusion in multimodal short video classification. Experimental results demonstrate that DMAFNet consistently outperforms other two mainstream methods in multimodal short video classification.

Although DMAFNet has achieved good results in short video classification, it faces challenges in efficiently achieving accurate classification when dealing with longer videos. To improve the efficiency of multimodal video classification, it is necessary to further explore how to reduce the amount of calculation.

References

[1] LAPTEV I. On space-time interest points[J]. *International Journal of Computer Vision*, 2005, 64(2): 107-123.

[2] WANG H, KLÄSER A, SCHMID C, et al. Dense trajectories and motion boundary descriptors for action recognition [J]. *International Journal of Computer Vision*, 2013, 103(1): 60-79.

[3] DONAHUE J, HENDRICKS L A, GUADARRAMA S, et al. Long-term recurrent convolutional networks for visual recognition and description [C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2015: 2625-2634.

[4] JI S, XU W, YANG M, et al. 3D convolutional neural networks for human action recognition [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012, 35(1): 221-231.

[5] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. California: ACM, 2017: 6000-6010.

[6] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks [C]//2015 IEEE International Conference on Computer Vision (ICCV). New York: IEEE, 2015: 4489-4497.

[7] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions [C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2015: 1-9.

[8] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? A new model and the kinetics dataset[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2017: 4724-4733.

[9] XIE S N, SUN C, HUANG J, et al. Rethinking spatiotemporal feature learning: speed-accuracy trade-offs in video classification [C]//15th European Conference on Computer Vision (ECCV). Berlin: Springer, 2018: 318-335.

[10] TRAN D, WANG H, TORRESANI L, et al. A closer look at spatiotemporal convolutions for action recognition [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 6450-6459.

[11] ARNAB A, DEHGHANI M, HEIGOLD G, et al. ViViT: a video vision transformer[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). New York: IEEE, 2021: 6816-6826.

[12] BERTASIUS G, WANG H, TORRESANI L. Is space-time attention all you need for video understanding? [C]//International Conference on Machine Learning (ICML). San Diego: JMLR, 2021:139.
[13] LI K C, WANG Y L, GAO P, et al. UniFormer: unified transformer for efficient spatiotemporal representation learning[EB/OL]. (2022-02-08)[2023-10-20]. https://arxiv.org/abs/2201.04676.
[14] DAVE I R, RIZVE M N, CHEN C, et al. TimeBalance: temporally-invariant and temporally-distinctive video representations for semi-supervised action recognition [C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2023: 2341-2352.
[15] HE B, LI H D, JANG Y K, et al. MA-LMM: memory-augmented large multimodal model for long-term video understanding[EB/OL]. (2024-04-24)[2024-05-16]. https://arxiv.org/abs/2404.05726.
[16] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [C]// Conference of the North-American-Chapter of the Association-for-Computational-Linguistics-Human Language Technologies (NAACL-HLT). Stroudsburg: ACL, 2019: 4171-4186.
[17] LIN R C, XIAO J, FAN J P. NeXtVLAD: an efficient neural network to aggregate frame-level features for large-scale video classification[C]// 15th European Conference on Computer Vision (ECCV). Berlin: Springer, 2018: 206-218.
[18] SUN C, MYERS A, VONDRICK C, et al. VideoBERT: a joint model for video and language representation learning [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). New York: IEEE, 2019: 7463-7472.

基于双流多模态对齐和融合的短视频分类网络

周 明[1,2], 王 彤[1,2*]
1. 东华大学 信息科学与技术学院, 上海 201620
2. 东华大学 数字化纺织服装技术教育部工程研究中心, 上海 201620

摘 要: 视频分类是视频理解中的一项重要任务, 在信息内容的智能监控中发挥着举足轻重的作用。大多数现有方法没有考虑视频的多模态性质, 而且模态融合方法往往过于简单, 常常忽略融合前的模态对齐。该文提出了一种用于短视频分类的双流多模态对齐和融合网络 DMAFNet。该网络使用两个单模态编码器来提取模态内的特征, 并且利用多模态编码器学习模态之间的交互。为了解决模态对齐问题, 引入了两个单模态编码器之间的对比学习。此外, 还设计了文本掩码建模和视频文本匹配辅助任务, 通过损失函数的反向传播来改善视频帧和文本模态之间的交互。实验证明了 DMAFNet 在多模态视频分类任务中的有效性。与两种主流的方法相比, DMAFNet 在 2022 年微信大数据挑战数据集上取得了最好的结果。

关键词: 视频分类; 多模态融合; 特征对齐