Multi-channel Attentive Weighting of Visual Frames for Multimodal Video Classification

Yuqing Wang[1], Zhuang Qi[1], Xiangxian Li[1], Jinxing Liu[1], Xiangxu Meng[1] and Lei Meng*[1,2]
[1] School of Software, Shandong University, Jinan, China
[2] Shandong Research Institute of Industrial Technology, Jinan, China
Email: {yuqing_wang, z_qi, xiangxian_lee, liujinxing}@mail.sdu.edu.cn, {mxx, lmeng}@sdu.edu.cn

***Abstract*—Multimodal video classification aims to incorporate semantic information to regularize the visual representation learning of videos. Conventional methods typically focus on analyzing all information extracted from different modals rather than key information. However, they usually face the problem of handling the redundant video frames of little categorical information. To address this problem, this paper proposes a novel approach that employs multi-channel weighting of visual frames to mitigate the interference of redundant information. Specifically, the proposed algorithm, termed MCA-WF, includes two main modules, where the multi-channel attentive weighting of video frames (McAW) module performs the multi-granularity and multi-channel frame weighting mechanism based on visual self-attention, contrastive attention and cross-modal attention constraints to filter visual noise and redundant information. The visual frame selection (VFS) module explores the combination of multi-channel attention mechanisms to select the key visual information in the video. Experiments were conducted on MSR-VTT and ActivityNet Captions datasets in terms of performance comparison, ablation study, in-depth analysis, and case studies. The results verified that MCA-WF can notice the key information in the classification and effectively improve the ability of information complementation and integration between modals, which leads to better performance than the state-of-the-art methods.**

***Index Terms*—Video classification, Multimodal information, Multi-channel, Key-frame selection, Attention.**

I. Introduction

Online video platforms have become one of the most popular applications that can collect short videos uploaded by users and also provide users with search functions based on video classes. Video classification has to deal with visual frames, text descriptions and audio information. Existing studies focus on extracting all modal information from the video. However, it has been observed that even the state-of-the-art methods [1]–[3] face problems in distinguishing the video classes in the diversity of video content scenarios due to the interference of redundant information and visual noise. Therefore, robust visual and semantic feature learning methods are urgently needed to extract key information.

The video classification task can be simply described as assigning the correct labels to visual content at the video level or the frame level. It is inherently continuous and multimodal, so deep neural models need to capture and aggregate the most relevant signals for a given input video. To achieve this goal, existing multimodal video classification methods typically follow two main approaches, which can be classified based on multimodal fusion methods [4], [5] and based on multi-image frame processing methods [6]–[9]. Methods based on multimodal fusion fuse information from different modalities to achieve information complementarity. Methods based on multi-frame processing extract and fuse video frames, using visual features in the video for classification. However, these methods lack the ability to select key frames in the video and to extract collaborative semantic information from the different modalities. This can lead to biases in feature fusion and model learning. At the same time, little thought has been given to the selection of video frames, resulting in a large number of redundant video frames that obscure the features between the different categories. Therefore, how to select high-quality keyframes and how to deal with the heterogeneity between modalities is an urgent problem that needs to be solved.

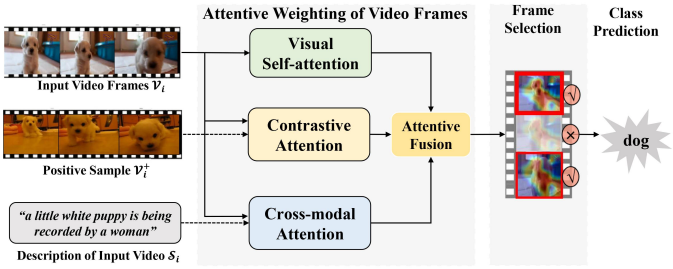


Fig. 1. Illustration of MCA-WF for video classification with multimodal information. The attentive weighting of video frames module fuses the multi-channel attention to generate weights of the video frames. Frame selection module selects the key frames of the video.

To address these issues, this paper presents a novel approach, termed MCA-WF, which is able to effectively mitigate the model fit deviation on training data caused by the lack of key information constraints, and enhance the representation learning ability between heterogeneous modalities. As illustrated in Figure 1, MCA-WF uses the **m**ulti-**c**hannel **a**ttentive **w**eighting of video **f**rames method to extract key frames while mitigating the influence of redundant and noisy video frames. Specifically, MCA-WF introduces a three-level weighting of visual frames method, including visual self-attention weighting (V-ATT), contrastive attention weighting (C-ATT), and cross-modal attention weighting (CM-ATT). In the V-ATT method, it is it possible to select important information in successive frames and filter out redundant frames that are not relevant

* indicates corresponding author.

to the video category. In the C-ATT method, content-level alignment of the video is achieved using the similarity between positive samples, so that background noise information can be filtered out and subject features relevant to the category can be emphasized. In the CM-ATT method, semantic information is leveraged to guide the selection of key visual information, enabling the selection of semantically relevant visual frames. Fusion of three-level weighting of visual frames based on multi-channel information to generate final keyframe selection weights. Attentive weights are used to select key visual frames to learn the final predicted class of the video.

Experiments are conducted on the MSR-VTT [10] and ActivityNet Captions [11] datasets in terms of performance comparison, ablation study of the key components of MCA-WF, and case studies for the effectiveness of three-level attentive weighting of video frames. The results verify that MCA-WF can improve the performance of different backbones, extract key visual information from the video and reduce the difference in the distribution of heterogeneous features. To summarize, this paper includes three main contributions:

- A general multimodal video classification algorithm MCA-WF based on a multi-channel attentive weighting mechanism of visual frames is proposed, which can effectively remove category-irrelevant information, achieve adaptive visual keyframe selection, and improve the accuracy of video classification.
- The contrastive attention weighting (C-ATT) method is proposed, which uses the similarity of features between positive samples to learn invariant representations with class-level information and identify key frames.
- McAW module can focus on the key information between different modalities. At the same time, it can learn uniform representations and reduce differences in the distribution of different modalities in the feature space.

II. Related Work

With the advancement of deep learning [12]–[17], video analysis techniques are increasingly being used in the fields of recommendation [18], [19], computer vision [20]–[27], etc. As video is a natural multimodal information source, multimodal learning [28]–[33] is essential for its analysis. Existing studies on video classification can be divided into two categories: methods based on single-modal information and methods based on multi-modal information.

A. Single-modal Information based Methods

Conventional video classification algorithms typically rely on single-modal visual information for video classification. The research content of methods based on single-modal information is the recognition of human behavior in simple scenes. There are two tasks. Behavior localization and behavior recognition. Behavior localization is the process of locating a video clip or frame with a target behaviour [34], [35]. Behavior recognition is to classify the behaviors in video clips [36], [37]. The prediction results are strongly affected by the redundant information of video frames in complex scenes. Therefore, complex scene video classification is more challenging than the original problem. With the rise of deep learning, researchers have started to use CNNs for video problems. The pioneering work DeepVideo [38] proposed to use a single 2D CNN model independently on each video frame and investigated several temporal connectivity patterns to learn spatio-temporal features for video action recognition, such as late fusion, early fusion and slow fusion.

B. Multi-modal Information based Methods

Existing multi-modal information methods use Efficientnet [39] and Nextvlad [40] to extract video, title and audio features respectively, concatenate multimodal features and use a neural network to learn fused features. Two-stream network [41] is a pioneering work in video understanding, which included a spatial stream and a temporal stream. Long et al. [42] proposed a multimodal keyless attention fusion for video classification. Neural machine translation [43] pointed out the video and audio features for video classification. At the same time, they introduced a context gating layer, an effective nonlinear unit called context gating, which was used to model the interdependence between network activations and showed stronger classification performance than LSTM [44] and GRU [3]. RNN is tried multimodal fusion and natural language processing technologies, such as attention mechanism [45], and it proposed a method of deep multimodal learning (DML), which combines visual and audio information at the video frame level. The DML model was tested in the Kaggle video classification contest. It shows good classification performance on large video datasets. However, these multimodal video classification algorithms lack key frame selection and can not filter out low-quality frames.

III. Method

In this section, we will detail the multimodal feature extraction (MFE) module, the multi-channel attentive weighting of visual frames (McAW) module, the visual frame selection (VFS) module, and the multimodal fusion network (MFN) module in MCA-WF and strategies to train the framework. The overview of MCA-WF is shown in Figure 2.

A. MFE Module for Visual and Semantic Features Extraction

Given a video dataset $\mathcal{D} = \{(\mathcal{V}_i, \mathcal{S}_i)|i = 1, ..., N\}$, $\mathcal{V}_i$ and $\mathcal{S}_i$ represent a video sample and the corresponding semantic information, respectively, where $\mathcal{V}_i = \{f_j \mid j = 1, \ldots, n\}$, f_j represents the frame of the input video, n is the number of frames, and $\mathcal{V}_i^+$ is a positive sample of $\mathcal{V}_i$. This semantic information can be audio or text descriptions. MFE module utilizes two encoders $\xi_v(\cdot;\theta_v)$ and $\xi_s(\cdot;\theta_s)$ to extract features from visual and semantic information respectively, where θ_v and θ_s denote the weights of visual and semantic encoders. The specific definitions are as follows:

$$\mathbf{F}_{v_i} = \xi_v(\mathcal{V}_i;\theta_v) \tag{1}$$

$$\mathbf{F}_{v_i}^+ = \xi_v(\mathcal{V}_i^+;\theta_v) \tag{2}$$

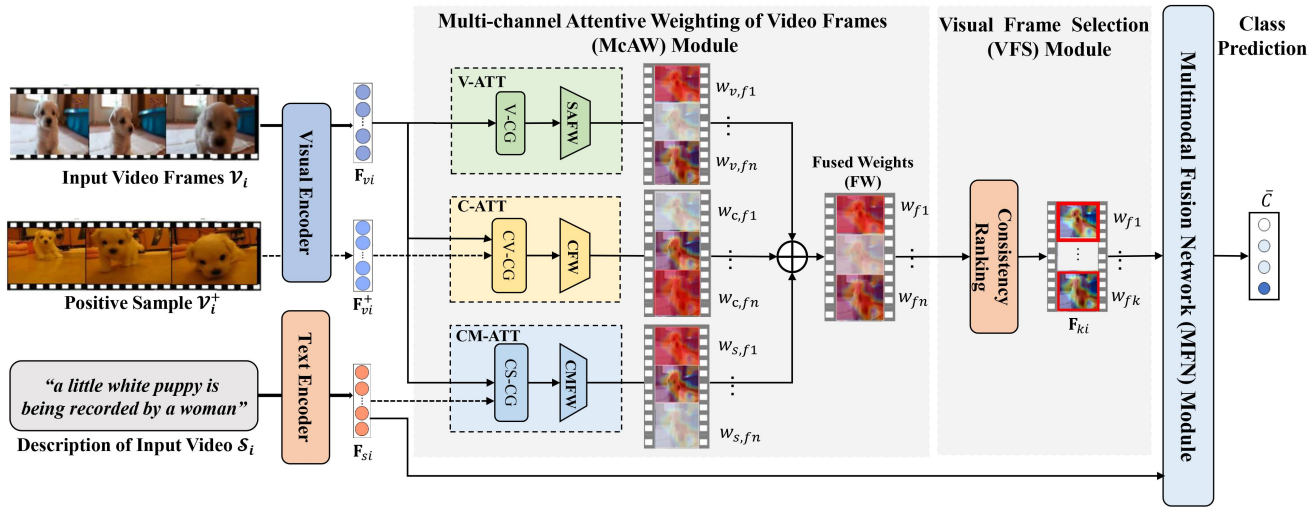


Fig. 2. Illustration of the proposed MCA-WF. Input the video frames information and the semantic information of the video, while **finding a positive sample for this video based on the same label.** The McAW module fuses the three levels of attentive weighting to obtain the fused weights of the visual frames. And it directs the VFS module to obtain the key frame of the video. Finally, the visual key frame information and text information are fused to the MFN module to predict the video label $\bar{C}$.

$$\mathbf{F}_{s_i} = \xi_s(\mathcal{S}_i; \theta_s) \tag{3}$$

This can extract the visual representation $\mathbf{F}_{v_i}$ and $\mathbf{F}_{v_i}^+$ with motion information from the original video $\mathcal{V}_i$ and the positive sample of the video $\mathcal{V}_i^+$, as well as the semantic representation $\mathbf{F}_{s_i} = \{\mathbf{s}_j \mid j = 1, \ldots, n\}$ of the original video with semantic information $\mathcal{S}_i$. Among them, $\mathbf{F}_{v_i} = \{\mathbf{v}_j \mid j = 1, \ldots, n\}$ and $\mathbf{F}_{v_i}^+ = \{\mathbf{v}_j^+ \mid j = 1, \ldots, n\}$, where $\mathbf{v}_j$ and $\mathbf{v}_j^+$represent the motion information of each frame in the video.

B. McAW Module for Visual Frames Weighting

In order to filter the noise and mine the key information of the video, the McAW module obtains the attentive weights from the three channels, including cross-modal attention weighting (CM-ATT), visual self-attention weighting (V-ATT), and contrastive attention weighting (C-ATT). The final weights of the visual frames are obtained by performing the fused weights (FW) to fuse the weights of the three parts of the video frames in order to select the key frames.

1) C-ATT for Content Level Alignment of the Video: **For each video, a positive video sample is selected based on the principle of similarity between the features of video frames with the same label.** The C-ATT is used to extract key information from the video and reduce intra-class differences using a similarity constraint between the positive sample video frames and the original video frames.

In C-ATT, positive sample visual frame features $\mathbf{F}_{v_i}^+$ inputs CV-CG to generate context vector c_c. Then c_c and $\mathbf{F}_{v_i}$ input the contrastive frame weighting (CFW) to get w_c. The contrastive attention weight $w_c = \{w_{c,f_i} \mid i = 1, \ldots, n\}$ of the positive sample to the original sample.

To further constrain the similarity of two videos that are positive samples of each other, the $\mathcal{L}_{ml}$ loss is also used to constrain the corresponding video frames. The specific definitions are as follows:

$$c_c = \mathcal{G}_c\left(\mathbf{F}_{v_i}^+\right) \tag{4}$$

$$w_c = Softmax\left(\theta_v\left(\mathbf{F}_{v_i}, c_c\right)\right) \tag{5}$$

$$\mathcal{L}_{ml} = \mathcal{M}\left(\mathbf{F}_{v_i}, \mathbf{F}_{v_i}^+\right) \tag{6}$$

where $\mathcal{G}_c(\cdot)$ represents the context vector generator. $\theta_v(\cdot,\cdot)$ represents fusion operation and linear mapping operation. $\mathcal{M}(\cdot,\cdot)$ represents linear mapping operation and the mean square error (MSE) function.

2) CM-ATT for Semantically Guided Visual Frames Selection: Semantic information is used to guide the weighting of visual frames, allowing semantically relevant visual frames and also initially reducing heterogeneity between modalities.

In CM-ATT, semantic information $\mathbf{F}_{s_i}$ inputs CS-CG to generate context vector c_s. Then c_s and $\mathbf{F}_{v_i}$ input the cross-modal frame weighting (CMFW) to obtain w_s. The cross-modal attention weight of each frame $w_s = \{w_{s,f_i} \mid i = 1, \ldots, n\}$ in the video is obtained. n indicates the number of frames in the video. The specific definitions are as follows:

$$c_s = \mathcal{G}_s\left(\delta\left(\mathbf{F}_{s_i}\right)\right) \tag{7}$$

$$w_s = Softmax\left(\theta_s\left(\mathbf{F}_{v_i}, \phi\left(c_s\right)\right)\right) \tag{8}$$

where $\delta(\cdot)$ is used to learn the deep semantic feature $\mathbf{F}_{sd}$. And then $\mathbf{F}_{sd}$ inputs to $\mathcal{G}_s(\cdot)$. CS-CG consists of $\delta(\cdot)$ and $\mathcal{G}_s(\cdot)$ operation. $\phi(\cdot)$ includes linear mapping and dimension expansion operations. $\theta_s(\cdot,\cdot)$ includes weighting operation.

3) V-ATT for Reducing the Redundancy of Visual Frames: The video frames are weighted by a self-attention visual weighting module that reduces redundant information in the continuous video frames and extracts key information. It can be used to further guide the weighting of the video frames.

In V-ATT, the video frame information $\mathbf{F}_{v_i}$ of the original video inputs V-CG to generate context vector c_v. Then c_v and $\mathbf{F}_{v_i}$ input self-attention frame weighting (SAFW) to obtain w_v. The weight w_v of each frame $w_v = \{w_{v,f_i} \mid i = 1, \ldots, n\}$ in

the video is obtained. n indicates the number of frames in the video. The specific definitions are as follows:

$$c_v = \mathcal{G}_v\left(\mathbf{F}_{v_i}\right) \tag{9}$$

$$w_v = Softmax\left(\theta_v\left(c_v, \mathbf{F}_{v_i}\right)\right) \tag{10}$$

where, $\mathcal{G}_v(\cdot)$ refers to the weighted average value of the visual frame features $\mathbf{F}_{v_i}$ through the context vector generator (V-CG). The c_v represents the features of the whole video frame.

4) FW for Multi-channel Weights Fusion of the Visual Frames : Fused the weights obtained from the three-level visual attention weighting modules. To better select the key information frames in the video, it can obtain the weights of the video frames from multiple perspectives.

In FW, the weights of multi-channel attention visual frames are fused. The input is the video frame weights w_s, w_v and w_c calculated by CM-ATT, V-ATT and C-ATT respectively. The output $w = \{w_{f_i} \mid i = 1, \ldots, n\}$ is the muti-channel attention video frame weight. The specific definitions are as follows:

$$w = Softmax(w_s + w_c + w_v) \tag{11}$$

C. VFS Module for Key Visual Frame Selection

In VFS module, the McAW module achieves potential alignment of semantic information on the content of visual information based on CM-ATT, filters redundant information between consecutive frames based on V-ATT, and reduces visual background noise based on C-ATT, respectively. Finally, the McAW achieves the constraint of key frames from different perspectives. Thus, this module can notice the key information in the classification after fusing the weights of the three parts visual frame weighting. Effectively improve the complementarity and integration of information between modalities.

The VFS module is used to select the key frame in the video. The multi-channel attention weights $w = \{w_{f_i} \mid i = 1, \ldots, n\}$ and video frames $\mathbf{F}_{v_i}$ as input to the consistency ranking (CR) module. The output is the key frame $\mathbf{F}_{k_i}$ in the video. The specific definitions are as follows:

$$\mathbf{F}, w_k = \Psi\left(\mathbf{F}_{v_i}, w\right) \tag{12}$$

$$\mathbf{F}_{k_i} = \mathbf{F} * w_k \tag{13}$$

where $\Psi(\cdot, \cdot)$ selects the top k weights of the video frames and obtains the original visual frames corresponding to the top k video frame weights as $\mathbf{F}$ in the video. The weight is weighted by the attention of multiple perspectives $w_k = \{w_{f_i} \mid i = 1, \ldots, k\}$, and finally take the dot product of the two to get the key frames in the video.

D. MFN Module for Multimodal Feature Fusion

Joint image-text representation learning methods mostly use dual-stream architectures to align image representations and text representations at the global level by contrast learning methods. However, it is easy to ignore fine-grained information to achieve effective alignment [46] and difficult to learn a representation that captures modality-invariant instance information corresponding to coherent natural language concepts. Therefore, in this paper, we use a single-stream Transformer architecture to achieve fine-grained alignment of visual and textual representations based on multi-headed attention and generate multimodal fusion representations.

The MFN module consists of a series of stacked Transformer blocks, including a multi-head self-attention layer (MSA) and MLP layer, which learn inter-modal attention encoded information based on a pre-trained multimodal network to fuse visual and semantic representations and generate multimodal fusion features. The specific formulas are as follows:

$$z^0 = [\mathbf{F}_{k_i}; \mathbf{F}_s] \tag{14}$$

$$\hat{e}^d = MSA\left(LN\left(e^{d-1}\right)\right) + e^{d-1}, \quad d = 1 \ldots D \tag{15}$$

$$e^d = MLP\left(LN\left(\hat{e}^d\right)\right) + \hat{e}^d, \quad d = 1 \ldots D \tag{16}$$

$$p = MLP\left(e^D\right) \tag{17}$$

where the Transformer layer number D is 12. In this module, the text embedding $\mathbf{F}_s$ and the multi-channel attentive weighting visual keyframe features $\mathbf{F}_{k_i}$ are concatenated into a multimodal fusion sequence e^0. Fusion features $\hat{e}^d$ are generated by MSA through implicit alignment of heterogeneous modal features and mapped to higher dimensional spaces to extract high-level abstract information. The prediction result p is generated by an MLP containing a two-layer linear mapping and an activation function *Relu(·)*, and the classification loss in the single label classification task is computed by the CrossEntropy (CE) loss $\mathcal{L}_{ce}$ with the true label.

E. Training Strategy of MCA-WF

The MCA-WF adopts a single-stage training mode to complete the migration fine-tuning of downstream tasks of the model through training. The model is updated by minimizing the weighted sum of predicted loss $\mathcal{L}_{ce}$ and MSE loss $\mathcal{L}_{ml}$. The specific formula of final loss is:

$$\mathcal{L}_{ce} = CE(p, p') \tag{18}$$

$$\mathcal{L}_{final} = \mathcal{L}_{ce} + \mathcal{L}_{ml} \tag{19}$$

where CE is the CrossEntropy function. p and p' are the predictions and ground truth of the label. $\mathcal{L}_{ml}$ is utilized according to (6).

IV. Experiments

A. Experiment Settings

1) Datasets: To demonstrate the generality of MCA-WF, we use two benchmarking datasets MSR-VTT and ActivityNet Captions that are commonly used in video classification for experiments. Their statistics are shown in Table I.

TABLE I
STATISTICS OF THE DATASETS USED IN THE EXPERIMENTS.

Datasets	#Class	#Sentences	#Training	#Testing
MSR-VTT	20	200,000	7,010	2,990
ActivityNet Captions	200	100,000	10,009	4,515

TABLE II
PERFORMANCE COMPARISON OF ALGORITHMS. METRICS ARE TOP-**1**/TOP-**5** ACCURARY (**ACC**). THE BEST PERFORMANCE IS MARKED IN BOLD.

Method	**Modality**	**MSR-VTT**		**ActivityNet Captions**	
		Acc@1%	**Acc@5%**	**Acc@1%**	**Acc@5%**
ViT (video)	V	53.7	82.9	81.9	94.1
GRU	V	49.5	79.2	78.6	93.8
MCA-WF(GRU)	V	53.8	83.8	82.3	94.8
GRU	V+S	53.1	81.8	80.3	94.0
MCA-WF(GRU)	V+S	56.4	84.2	83.9	95.3
ViLT	V+S	55.4	83.9	82.7	95.2
MCA-WF(ViLT)	V+S	**58.8**	**85.3**	**85.6**	**96.5**

- MSR-VTT dataset: contains 10,000 unique YouTube video clips. Each of them is annotated with 20 different text captions, so there are 200,000 video caption pairs in total. We split the dataset into 7,010 and 2,990 videos for training and testing, respectively.
- ActivityNet Captions dataset: contains 20,000 subtitled videos, each with a unique start and end time, totaling 849 hours of video, with 100,000 segments. On average, each 20,000 video contains 3.65 temporally localized sentences, for a total of 100,000 sentences. Since some videos are not officially labeled, the labeled sample data is divided into 10,009 and 4,515 videos for training and testing, respectively.

2) Evaluation Criteria: In the experiments on the MSR-VTT and ActivityNet Captions datasets, we use Accuracy to evaluate the model prediction performance in single-label classification, the accurary formula is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{20}$$

where TP is the number of positive samples, FP is the number of negative samples, TN is the number of positive samples and FN is the number of negative samples. We followed conventional measures of Top-1 and -5 accuracies to evaluate the classification performance. To alleviate the problem of randomness, we repeat the evaluation process five times and report the average value.

3) Implementation Details: To verify the applicability of MCA-WF, we investigate the performance of MCA-WF on three visual backbones ViLT [2] and GRU [3], denoted as MCA-WF (ViLT) and MCA-WF (GRU). The feature dimension is set to 768 in the experiment. This follows the feature dimension setting of the pre-trained large model ViLT. The batch size was selected from $\{32, 64, 128\}$. We used the Adam optimizer with a learning rate chosen from 1e-6 to 1e-3. The decay rate of the learning rate parameter was chosen from 0.1 and 0.5 with a decay interval of 4 epochs. We conduct experiments on NVIDIA Tesla V100 and a ViLT-based model takes 5$\sim$8 hours to train.

We used the pre-trained S3D network to extract visual features with a feature dimension of 1024 from two multimodal video datasets: MSR-VTT and ActivityNet Captions. To extract text features from sentence descriptions in MSR-VTT, we used the Google Cloud Speech to Text API network with a dimension of 768 and a sequence length of 73 [**?**]. For both datasets, we extracted 30 visual frames with the same interval. For the ActivityNet Captions dataset, we used the VGGish network, pre-trained on the YouTube-8M dataset, to extract audio features with a dimension of 128 from the semantics [47]. We also extracted 60 visual frames and 60 audio frames from the video using the same interval.

B. Performance Comparison

In this section, we show the effect of the MCA-WF on two multimodal datasets and compare the Vision Transformer (ViT) [1], Vision and Language Transformer (ViLT) [2] and GRU models as SOTAs experiments. On this basis, our proposed multimodal video classification module MCA-WF is merged. Our modules are also added to the GRU as MCA-WF(GRU) and added to ViLT as MCA-WF (ViLT). The following observations can be drawn from Table II.

- Our study shows that video classification performance improves when multimodal information is fused, compared to classification based on single-modal visual features. This suggests that there is complementary semantic information between different modes when feature fusion is used. Moreover, the fusion of semantic and visual information yields the most significant improvement in classification accuracy.
- Using a multi-channel attentive weighting of visual frames leads to improved multimodal video classification. This method, which includes CM-ATT, V-ATT, and C-ATT, enables the selection of key information frames from various perspectives. Specifically, CM-ATT combines semantic and visual information for collaborative learning, V-ATT reduces redundant information and filters out visual frame noise, and C-ATT eliminates background information and emphasizes body-related information for the given category.
- The overall effect of the pre-trained large model ViLT on the experimental dataset is higher than that of the backbone GRU network, indicating the superiority of the pre-trained large model in downstream tasks. At the same time, the MFN module is able to further reduce the deviation that is caused by the uneven distribution between the different modalities.

TABLE III
ABLATION STUDY OF MCA-WF WITH **VILT** BACKBONE IN TERMS OF TOP-**1** AND TOP-**5** ACCURACY (**ACC**). (V_{attv} + S),(V_{atts} + S) AND (V_{attc} + S) DENOTE THE VISUAL FUSION SEMANTIC INFORMATION AFTER THE V-ATT, THE CM-ATT, AND THE C-ATT METHODS, RESPECTIVELY. THE BEST PERFORMANCE IS MARKED IN BOLD.

Method	MSR-VTT		ActivityNet Captions	
	Acc@1%	Acc@5%	Acc@1%	Acc@5%
V	51.3	81.6	82.0	94.1
V + S	55.4	83.9	82.7	95.2
V_{atts} + S	57.5	84.1	85.1	96.3
V_{attv} + S	57.9	84.2	85.0	96.2
V_{attc} + S	57.6	84.3	85.1	96.4
$V_{atts+attv}$ + S	58.0	84.7	85.4	96.3
$V_{atts+attc}$ + S	58.1	84.5	85.2	96.3
$V_{attv+attc}$ + S	58.1	84.6	85.3	96.4
$V_{atts+attv+attc}$ + S	**58.8**	**85.3**	**85.6**	**96.5**

- After applying the MCA-WF multimodal video classification algorithm to different models (GRU or ViLT), the video classification performance is significantly improved, demonstrating its model-independent properties.

C. Ablation Study

In this section, we further studied the working mechanisms of different modules of MCA-WF, as shown in Table III. We chose ViLT as the base network. The following findings could be observed:

- Use multimodal information fusion (V+S) outperforms single-modal visual features (V). This is due to the consistency between multimodal information and the complementarity between fine-grained modalities.
- The performance of the model is improved by all three parts of the visual frame weighting. (V_{atts} + S) achieves a potential alignment between the visual and the semantic information of the video. (V_{attv} + S) eliminates redundant information from the video frames. The (V_{attc} + S) module achieves content alignment between video feature frames and category-related content, highlighting the subject information.
- Pairwise combination of the weighting mechanism of the three perspectives ($V_{atts+attv}$ + S), ($V_{attv+attv}$ + S) and ($V_{atts+attc}$ + S) can further improve the performance of the video classification. The ($V_{atts+attv+attc}$ + S) module achieves the best performance on two realistic multimodal video datasets. It shows that the attention module of the three parts with different perspectives can effectively constrain the key frames of the video.

D. In-depth Analysis

1) Evaluation on Frame Weighting for Visual Modal: The CM-ATT, V-ATT and C-ATT are the main modules proposed in this paper. On the one hand, the modal is different; on the other hand, the visual frames are weighted from different perspectives. Therefore, we further analyze the effectiveness of the three video frame weighting methods without using semantic information. Experimental results show that all three parts of attentional weighting can constrain video keyframes. The result is shown in Table IV.

TABLE IV
THE PERFORMANCE COMPARISON OF THE THREE PARTS OF ATTENTION FRAME WEIGHTING IN THE VISUAL MODAL IN TERMS OF TOP-**1** AND TOP-**5** ACCURACY (**ACC**). EXPERIMENTS ARE VALIDATED ON THE BASE MODEL **VILT**. THE BEST PERFORMANCE IS MARKED IN BOLD.

Method	MSR-VTT		ActivityNet Captions	
	Acc@1%	Acc@5%	Acc@1%	Acc@5%
V	51.3	81.6	82.0	94.1
V_{attv}	53.6	82.7	83.5	95.8
V_{atts}	**53.9**	**83.0**	83.6	95.7
V_{attc}	53.5	82.6	**83.9**	**96.0**

TABLE V
COMPARE THE PERFORMANCE OF SEMANTIC INFORMATION, VIDEO INFORMATION, AND MULTIMODAL INFORMATION IN VIDEO CLASSIFICATION IN TERMS OF TOP-**1** AND TOP-**5** ACCURACY (**ACC**). V: VISUAL INFORMATION. S: SEMANTIC INFORMATION. THE BEST PERFORMANCE IN DIFFERENT MODELS IS MARKED IN BOLD.

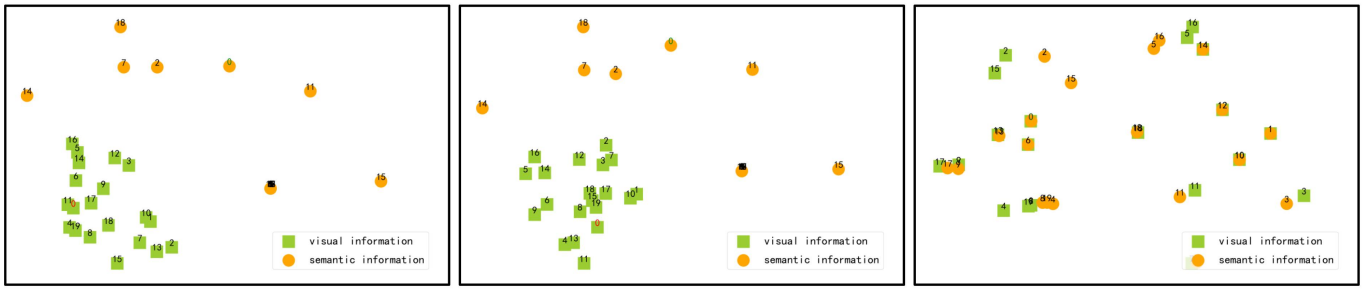
Backbone	Modality	MSR-VTT		ActivityNet Captions	
		Acc@1%	Acc@5%	Acc@1%	Acc@5%
GRU	V	49.5	79.2	78.6	93.8
	S	52.5	81.2	22.5	42.5
	V+S	**53.1**	**81.8**	**80.3**	**94.0**
ViLT	V	51.3	81.6	82.0	94.1
	S	53.0	81.5	24.8	45.8
	V+S	**55.4**	**83.9**	**82.7**	**95.2**

2) Evaluation on Different Modality for Video Classification Performance: Different modality of the video has different impacts on the video classification performance. Experiments are conducted on GRU and ViLT backbones. The result is shown in Table V. The following findings could be observed:

- The experiments show that the classification effect of multimodal video fusion is higher than single-mode video in both ViLT and GRU models. Illustrates the complementarity and consistency of two types of data features between different modal information.
- In the MSR-VTT dataset, semantic information is a textual feature, and the classification effect of semantic information is better than that of visual information. This indicates the importance of higher-order semantic information. However, in the ActivityNet Captions dataset, the semantic information is not as good as the visual information for the classification of the videos. **Because it uses audio features, audio features come with some noise information. But it can still have some classification effects.** The higher-order correlation information between vision and semantics can be better extracted by combining the two.

E. Case Studies

1) Quality Analysis of Representation Learning: This section analyses the video classification MCA-WF algorithm according to the visual and semantic representation distribution at different stages. Figure 3 shows the results. (a) shows



(a) Shallow representation (b) Multi-channel attentive representation (c) Interaction representation

Fig. 3. Visualization of visual and semantic information of 20 randomly selected test samples in MCA-WF module. "Orange" represents semantic information and "Green" represents visual information. (a) is the distribution of visual and semantic shallow representations obtained by the feature extraction (MFE) module; (b) is the distribution of visual and semantic representations after the multi-channel attentive frame weighting (McAW) mechanism; (c) represents visual and semantic distribution after the multimodal fusion network (MFN) module.

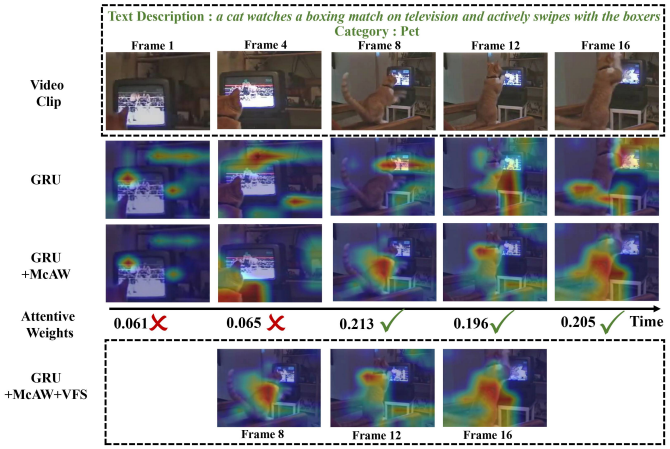


Fig. 4. Illustration of the results of the multi-channel attentive visual frame weighting (McAW) module and the visual frame selection (VFS) module for selecting key visual frames on the GRU model.

the shallow representations obtained in the MFE module; (b) shows the features of multi-channel attentive representations extracted in the McAW module; (c) shows multimodal interactive features extracted in the MFN module.

From Figure 3, it can be seen that the representation distribution of different modalities of visual and semantic information has obvious changes in the embedding space of t-SNE [48]. The distribution of visual and semantic representations of multimodal video is slowly converging.

- In Figure 3(a), the distribution of visual and semantic superficial representations of the same video shows an obvious distinction.
- In Figure 3(b), the distribution of visual and semantic information in feature space tends to be close, indicating that the attentional mechanism can achieve frame selection with deep semantics and key frame extraction.
- In Figure 3(c), the representations of the two modalities gradually become consistent, indicating that the MFN module can alleviate the problem of inconsistent distribution of heterogeneous features.

2) The Visual Verification of Frame-weighted Selection Performance: To verify the impact of the module McAW module and the VFS module on experimental performance, the corresponding case studies are performed on the GRU model. As shown in Figure 4, the first line of the video clip represents a moderately spaced selection of five visual frames. The video's label is Pet and the corresponding text description is 'a cat watches a boxing match on television and actively swipes with the boxers'. As the results show, the original GRU model focuses more on the subject of 'television', suggesting that the context interferes with the model's focus on the characteristics of the subject. In the 'GRU+McAW' method, the three-level attentive frames fusion gives different weights to different visual frames and reduces the influence of the background. The McAW allows the model to focus on the central part 'cat' of the video, removing background subjects 'television'. The McAW module enhances visual representation learning. In the VFS module, high-quality frames are selected and the redundant information is filtered out. It is illustrated that the two-part module can further extract key information from the video while filtering out redundant information from the video.

V. Conclusion

This paper proposes a multimodal video classification method based on multi-channel attentive weighting of video frames (MCA-WF). MCA-WF uses a multi-granularity and multi-channel frame weighting mechanism to filter visual noise and redundant information in the visual modal. This enhances the complementarity and integration ability of key information between different modals and can pay attention to the key information in the classification.

The MCA-WF algorithm effectively mitigates the complementarity of information between different modalities, extracts the key semantic information of each modality, and learns the unified representation. It can effectively constrain the key frames in the video to improve the accuracy of the video classification. In the next step, an attempt will be made to further establish an efficient key frame selection mechanism.

ACKNOWLEDGMENT

This work is supported in part by the Excellent Youth Scholars Program of Shandong Province (Grant no.2022HWYQ-048), the TaiShan Scholars Program (Grant no.tsqn202211289), and the National Key R&D Program of China (Grant no.2021YFC3300203).

REFERENCES

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[2] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *ICML*, 2021.

[3] A. Miech, I. Laptev, and J. Sivic, "Learnable pooling with context gating for video classification," *arXiv preprint arXiv:1706.06905*, 2017.

[4] C. Yan, Y. Tu, X. Wang, Y. Zhang, X. Hao, Y. Zhang, and Q. Dai, "Stat: Spatial-temporal attention mechanism for video captioning," *IEEE transactions on multimedia*, 2019.

[5] X. Yang, P. Molchanov, and J. Kautz, "Multilayer and multimodal fusion of deep neural networks for video classification," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016.

[6] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *Proceedings of ECCV*, 2016.

[7] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[8] T. Zhao, "Deep multimodal learning: An effective method for video classification," in *Proceedings of ICWS*, 2019.

[9] A. Graves, "Long short-term memory," *Supervised sequence labelling with recurrent neural networks*, 2012.

[10] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *Proceedings of CVPR*, 2016.

[11] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proceedings of CVPR*, 2015.

[12] Z. Wang, C. Li, and X. Wang, "Convolutional neural network pruning with structural redundancy reduction," in *Proceedings of CVPR*, 2021.

[13] L. Meng, A.-H. Tan, and D. Wunsch, *Adaptive resonance theory in social media data clustering*, 2019.

[14] Y. Wang, X. Li, Z. Qi, J. Li, X. Li, X. Meng, and L. Meng, "Meta-causal feature learning for out-of-distribution generalization," in *European Conference on Computer Vision*, 2023.

[15] L. Meng, A.-H. Tan, and D. C. Wunsch, "Adaptive scaling of cluster boundaries for large-scale social media data clustering," *IEEE transactions on neural networks and learning systems*, 2015.

[16] L. Meng and A.-H. Tan, "Semi-supervised hierarchical clustering for personalized web image organization," in *Proceedings of IJCNN*, 2012.

[17] Y. Wang, X. Li, H. Ma, Z. Qi, X. Meng, and L. Meng, "Causal inference with sample balancing for out-of-distribution detection in visual classification," in *Artificial Intelligence: Second CAAI International Conference, CICAI 2022, Beijing, China, August 27–28, 2022, Revised Selected Papers, Part I*, 2022.

[18] L. Meng, F. Feng, X. He, X. Gao, and T.-S. Chua, "Heterogeneous fusion of semantic and collaborative information for visually-aware food recommendation," in *Proceedings of ACM MM*, 2020.

[19] H. Ma, X. Li, L. Meng, and X. Meng, "Comparative study of adversarial training methods for cold-start recommendation," in *Proceedings of the 1st International Workshop on Adversarial Learning for Multimedia*, 2021.

[20] L. Meng, L. Chen, X. Yang, D. Tao, H. Zhang, C. Miao, and T.-S. Chua, "Learning using privileged information for food recognition," in *Proceedings of ACM MM*, 2019.

[21] L. Meng, Q. H. Nguyen, X. Tian, Z. Shen, E. S. Chng, F. Y. Guan, C. Miao, and C. Leung, "Towards age-friendly e-commerce through crowd-improved speech recognition, multimodal search, and personalized speech feedback," in *Proceedings of the 2nd International Conference on Crowd Science and Engineering*, 2017.

[22] X. Li, H. Ma, L. Meng, and X. Meng, "Comparative study of adversarial training methods for long-tailed classification," in *Proceedings of the 1st International Workshop on Adversarial Learning for Multimedia*, 2021.

[23] L. Wu, X. Chen, L. Meng, and X. Meng, "Multitask adversarial learning for chinese font style transfer," in *Proceedings of IJCNN*, 2020.

[24] X. Chen, L. Wu, M. He, L. Meng, and X. Meng, "Mlfont: Few-shot chinese font generation via deep meta-learning," in *Proceedings of the 2021 International Conference on Multimedia Retrieval*, 2021.

[25] P. Dong, L. Wu, L. Meng, and X. Meng, "Hr-prgan: High-resolution story visualization with progressive generative adversarial networks," *Information Sciences*, 2022.

[26] J. Li, H. Ma, X. Li, Z. Qi, L. Meng, and X. Meng, "Unsupervised contrastive masking for visual haze classification," in *Proceedings of ICMR*, 2022.

[27] W. Guo, Y. Zhang, X. Cai, L. Meng, J. Yang, and X. Yuan, "Ld-man: Layout-driven multimodal attention network for online news sentiment recognition," *IEEE Transactions on Multimedia*, 2020.

[28] L. Meng, A.-H. Tan, C. Leung, L. Nie, T.-S. Chua, and C. Miao, "Online multimodal co-indexing and retrieval of weakly labeled web image collections," in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, 2015.

[29] L. Meng, A.-H. Tan, and D. Xu, "Semi-supervised heterogeneous fusion for multimedia data co-clustering," *IEEE Transactions on Knowledge and Data Engineering*, 2013.

[30] J. Liu, J. Xiao, H. Ma, X. Li, Z. Qi, X. Meng, and L. Meng, "Prompt learning with cross-modal feature alignment for visual domain adaptation," in *Artificial Intelligence: Second CAAI International Conference, CICAI 2022, Beijing, China, August 27–28, 2022, Revised Selected Papers, Part I*, 2022.

[31] L. Meng and A.-H. Tan, "Community discovery in social networks via heterogeneous link association and fusion," in *Proceedings of the 2014 SIAM International Conference on Data Mining*, 2014.

[32] A.-H. Tan, B. Subagdja, D. Wang, and L. Meng, "Self-organizing neural networks for universal learning and multimodal memory encoding," *Neural Networks*, 2019.

[33] C. Lin, S. Zhao, L. Meng, and T.-S. Chua, "Multi-source domain adaptation for visual sentiment classification," in *AAAI*, 2020.

[34] A. Kläser, M. Marszałek, C. Schmid, and A. Zisserman, "Human focused action localization in video," in *Proceedings of ECCV*, 2010.

[35] Z. Shou, D. Wang, and S.-F. Chang, "Temporal action localization in untrimmed videos via multi-stage cnns," in *Proceedings of CVPR*, 2016.

[36] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep bi-directional lstm with cnn features," *IEEE access*, 2017.

[37] A. Alfaro, D. Mery, and A. Soto, "Action recognition in video using sparse coding and relative features," in *Proceedings of CVPR*, 2016.

[38] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of CVPR*, 2014.

[39] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, 2019.

[40] R. Lin, J. Xiao, and J. Fan, "Nextvlad: An efficient neural network to aggregate frame-level features for large-scale video classification," in *Proceedings of ECCV Workshops*, 2018.

[41] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in neural information processing systems*, 2014.

[42] X. Long, C. Gan, G. Melo, X. Liu, Y. Li, F. Li, and S. Wen, "Multimodal keyless attention fusion for video classification," in *AAAI*, 2018.

[43] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[44] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, "Video captioning with attention-based lstm and semantic consistency," *IEEE Transactions on Multimedia*, 2017.

[45] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, 2021.

[46] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proceedings of ICML*, 2021.

[47] V. Gabeur, C. Sun, K. Alahari, and C. Schmid, "Multi-modal Transformer for Video Retrieval," in *Proceedings of ECCV*, 2020.

[48] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, 2008.