Pattern Recognition 140 (2023) 109578

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Topic-aware video summarization using multimodal transformer

Yubo Zhu[a], Wentian Zhao[a], Rui Hua[a], Xinxiao Wu[a,b,*]

[a] Beijing Key Laboratory of Intelligent Information Technology, School of Computer Science & Technology, Beijing Institute of Technology, China
[b] Guangdong Laboratory of Machine Perception and Intelligent Computing, Shenzhen MSU-BIT University, China

ARTICLE INFO

Article history:
Received 15 June 2022
Revised 12 March 2023
Accepted 28 March 2023
Available online 30 March 2023

Keywords:
Topic-aware video summarization
Multimodal transformer
Video summarization dataset

ABSTRACT

Video summarization aims to generate a short and compact summary to represent the original video. Existing methods mainly focus on how to extract a general objective synopsis that precisely summaries the video content. However, in real scenarios, a video usually contains rich content with multiple topics and people may cast diverse interests on the visual contents even for the same video. In this paper, we propose a novel topic-aware video summarization task that generates multiple video summaries with different topics. To support the study of this new task, we first build a video benchmark dataset by collecting videos from various types of movies and annotate them with topic labels and frame-level importance scores. Then we propose a multimodal Transformer model for the topic-aware video summarization, which simultaneously predicts topic labels and generates topic-related summaries by adaptively fusing multimodal features extracted from the video. Experimental results show the effectiveness of our method.

1. Introduction

Automatic video summarization refers to generating a short synopsis that summarizes the video by exhibiting its most informative and important parts, so users can quickly grasp the main idea of a video without spending much time to watch the whole content. A wide range of applications can be benefit from it such as generating teasers of movies and episodes of a TV series, presenting the highlights of an event (e.g., a sports game, a music band performance, or a public debate) and improving video sharing platforms' viewing experience.

Early traditional methods [1] of video summarization cluster low-level visual features (e.g., appearance and motion features) to generate video summaries. Considerable progress has been made through extracting more expressive visual features through deep neural networks in recent deep learning methods [2,3]. All these methods focus on generating a single-perspective video summary to represent the overall video content without any personalized information.

It is the fact that video summarization is a subjective task, since there could coexist multiple topics such as animal, character and scenery in the same video and users may be attracted by different topics. For example, when facing with a celebrity cooking video, the food lovers may concern more about the process of making food, and the celebrity fans may be more interested in the person who is cooking. Moreover, take the "Life of Pi" movie as an example. This movie tells a story of the protagonist Pi and the Bengal tiger drifting on a boat, and different users may have different potentially favorite video clips of the same movie. The users who are animal lovers are eager to see Bengal tigers or other animals. The users who just want to know how the protagonist Pi survives in such a hard situation prefer a summary related to the protagonist.

Therefore, generating a single video summary from any objective perspective is far from enough to represent different topics of the same video, and cannot meet subjective needs with different preferences. In this paper, what we are eager to investigate is generating multiple video summaries on different topics under the premise of representing the main content of the video, so as to achieve the personalization of video summaries. With this in mind, we propose a topic-aware video summarization task, as shown in Fig. 1, which not only enables users to grasp the main content of the video in a short time, but also can generate video summaries on different topics to meet the interests of different users. Topic-aware video summarization takes into account the subjectivity of video summarization and allows users to have more choices, rather than generating a single summary of an input video in existing methods [2,3]. Topic-aware video summarization is also more in line with real applications, where video platforms often treat topics as labels to identify videos and recommend videos. In fact, some current online video platforms like YouTube are paying

* Corresponding author at: Beijing Key Laboratory of Intelligent Information Technology, School of Computer Science & Technology, Beijing Institute of Technology, China.
E-mail addresses: 3120211052@bit.edu.cn (Y. Zhu), wentian_zhao@bit.edu.cn (W. Zhao), huarui@bit.edu.cn (R. Hua), wuxinxiao@bit.edu.cn (X. Wu).

https://doi.org/10.1016/j.patcog.2023.109578
0031-3203/

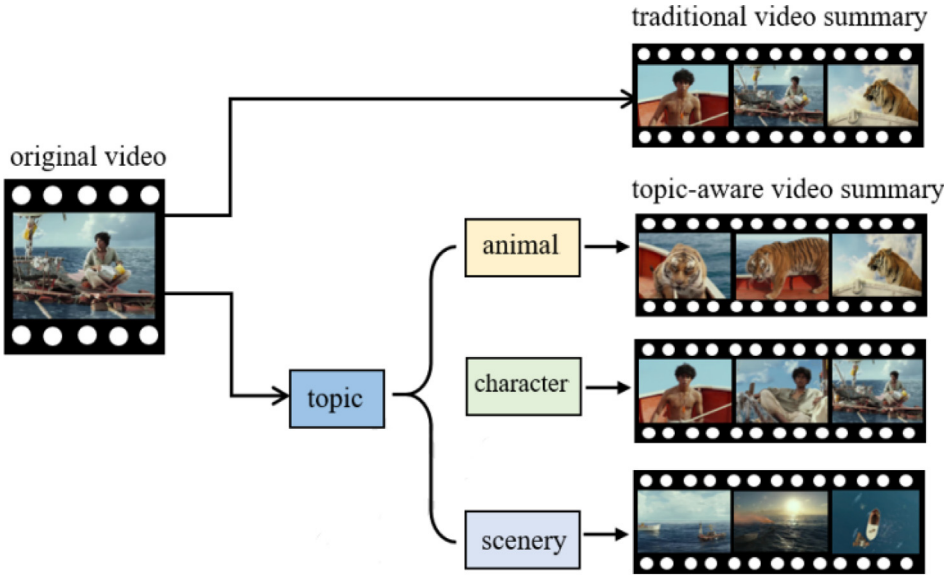

Fig. 1. Topic-aware video summarization. The traditional video summary can only extract a single summary from the original video, but topic-aware summary can extract multiple video summaries according to different topics.

attention to video topics. Video platforms will be able to grasp the types of videos that users concern according to the video topics that users frequently watch, so that personalized recommendations can be achieved. Generating video summaries conformed to the users' favorite topic can greatly increase the users' interest in the video, and can also increase the users' degree of dependence on the video platforms.

However, topic-aware video summarization remains challenging as it is extremely difficult to simultaneously predict topic labels and extract the informative video parts. To support the study of this new task, we build a new video dataset, called TopicSum. It contains 136 five-minutes-long videos with larger scale, richer data and more diversification compared with existing video summarization datasets, such as TVSum [4] and SumMe [5]. In order to support the topic-aware video summarization, our TopicSum dataset not only has the annotations of frame-level importance scores, but also provides the annotations of topic labels. In addition, TopicSum contains three different modalities of visual, textual and audio sampled from videos rather than a single visual modality in other existing datasets to support the multimodal task.

We also propose a topic-aware video summarization method based on a multimodal Transformer to meet the challenge of generating high-quality video summaries of different topics. Our method jointly models the prediction of importance scores and the classification of topics, selects representative shots according to the importance scores, satisfies the interests of different users through topic classification, and finally generates topic-aware video summaries. In order to make full use of the rich information in videos, we adaptively fuse the visual feature, the audio feature and the textual feature extracted from video via the multimodal Transformer. For complex scenes in videos, it is difficult for a single modality to provide sufficient information, which may lead to incorrect prediction results. Synthesising the knowledge of multiple modalities makes up for the deficiency of single modality, and thus benefits making further decisions and improving the quality of summaries.

Extensive experiments demonstrate that our method succeeds in generating topic-aware video summaries for different users to pay attention to the topics they are interested in. In summary, the contributions of our work are:

- We propose a novel topic-aware video summarization task that focuses on generating multiple topic-related video summaries, which especially advantageous in meeting the subjective needs of different users.
- We contribute a topic-aware video summarization dataset, called TopicSum, with a view to increasing attention to the impact of user interests on video summarization.
- We propose a novel multimodal Transformer model which can adaptively fuse multimodal features to make up for the deficiency of single modality, thus benefits making further decisions and improving the quality of summaries.

2. Related work

2.1. Video summarization

Early traditional methods of video summarization generally extract the handcrafted visual feature such as color histogram [6], spatio-temporal feature [7] and motion cues [8], and use clustering methods to generate video summaries.

Due to the great success of deep learning in video processing and understanding, many methods based on deep neural networks have been proposed. Zhang et al. [9] formulate the video summarization as a selection process of a subset of video shots, and implement the generation of video summaries by using Long Short-Term Memory (LSTM) units. Zhao et al. [10] improve the LSTM network using a layered structured adaptive network to extract video summaries. Since the recurrent neural networks are not suitable for processing the complex structure of long videos, Rochan et al. [11] use a fully convolutional model instead of recurrent neural networks to evaluate and select video frames for generating high-quality video summaries. Considering that the predicted scores of video frames in the same semantic segment cannot accurately represent the importance of the corresponding segment, Zhu et al. [12] attempt to use temporal consistency via the temporal interest detection formulation to determine and localize the representative contents of video sequences. More recently, the attention mechanism has been employed in video summarization. Ghauri et al. [13] suggest a novel model architecture that combines three feature sets for visual content and motion to predict impor-

Table 1
Comparison between our method and some recent methods from the perspectives of whether they have supervised information, employ attention mechanisms, use multimodal feature (visual, textual, audio), and generate topic-aware video summaries.

Method	Supervised	Attention	Visual	Textual	Audio	Topic-aware
Hsa-rnn [10]	✓		✓			
GDPP [17]	✓		✓			
SUM-FCN [11]	✓		✓			
DSNet [12]	✓	✓	✓			
MSVA [13]	✓	✓	✓			
Zhu et al. [2]	✓	✓	✓			
AC-SUM-GAN [14]			✓			
3DST-UNet [15]			✓			
DSAVS [16]		✓	✓	✓		
SASUM [18]	✓		✓	✓		
CHAN [19]	✓		✓	✓		
CLIP-It [20]	✓	✓	✓	✓		
Ours	✓	✓	✓	✓	✓	✓

tance scores. They utilize an attention mechanism before fusing motion features and features representing the visual content. Zhu et al. [2] propose a multiscale hierarchical attention approach that exploits the underlying hierarchical structure of video sequences and learns both the short-range and long-range temporal representations via a intra-block and a inter-block attention.

Unsupervised video summarization attracts growing attentions and has seen considerable progress through generative adversarial networks and reinforcement learning models. Apostolidis et al. [14] embed an actor-critic model into a generative adversarial network and formulate the selection of important video fragments as a sequence generation task. Liu et al. [15] implement unsupervised video summarization with reinforcement learning, and a 3D spatio-temporal U-Net is used to efficiently encode spatio-temporal information of the input videos for downstream reinforcement learning. Zhong et al. [16] propose a deep semantic and attentive network for Video Summarization (DSAVS) that generates unsupervised video summary by minimizing the distance between video representation and text representation, and introduce a self-attention mechanism to capture the long-range temporal dependencies.

Different from the aforementioned methods that generate a single and general video summary without considering user preferences on the extracted video parts, our method generates multiple video summaries on different topics and thus can meet various interests of users.

2.2. Multimodal feature learning

Since there exist rich information in videos such as vision, text and audio, multimodal features learning has also been studied in video summarization. Yuan et al. [21] extract semantic information from the side, including video titles, user query, video description and user comments, and define a video summary by maximizing the relevance between visual and semantic features in a common latent space. Wei et al. [18] use manual description annotations for videos and select video shots by minimizing the distance between the generated description sentence of the summary and the human annotated text of the original video, with the help of semantic attended networks. Query-focused video summarization is an application of multimodal feature learning. Xiao et al. [19] formulate the task as a problem of computing similarity between video shots and query, and propose a convolutional network with local self-attention mechanism and query-aware global attention mechanism to learn visual information of each shot. Li et al. [22] apply multimodal feature learning to self-supervised video summarization. They explore the semantic consistency between the videos and text in both coarse-grained and fine-grained fashions, as well as recovering masked frames in the videos.

The most related to our method is [20], which proposes a language-guided multimodal transformer that learns to score frames in a video based on their importance relative to one another and their correlation with a user-defined query (for query-focused summarization) or an automatically generated dense video caption (for generic video summarization). Rather than requiring a language query input, our method simultaneously generates multiple video summaries with different topics, which can cater different users' interests. Moreover, most existing methods use the semantic information to guide the training of the summary generation model. In contrast, our multimodal Transformer adaptively fuses the visual, textual and audio features, succeeding in the topic-aware video summarization.

Table 1 shows the comparison between our method and some recent methods mentioned above.

3. Dataset

3.1. Video collection and annotation

We build a video dataset of topic-aware video summarization, named TopicSum, that consists of 136 content-rich videos sampled from various movies such as "Life of Pi" and "The Chronicles of Narnia". The types of movie videos include but are not limited to comedy, family, and biography. We sample five-minutes-long videos from the main part of each movie. Table 2 shows more detailed information about TopicSum, including the name of source movies, the resolution of video frames, the sampling duration, and the number of videos sampled from each movie. We remove the opening titles and the closing credits from each movie. The subtitles are used as the text information corresponding to each video.

Table 2
Details of the TopicSum dataset, including the name of source movies, the resolution of video frames, the sampling duration, and the number of videos sampled from each movie.

Source Movie	Resolution	Duration(h)	Video Num.
Eight Below	1920 × 800	1.92	23
A Dog's Purpose	1280 × 536	1.50	18
Jane(2017)	1920 × 1072	1.42	17
The Chronicles of Narnia	1920 × 816	1.67	20
Life of Pi	1920 × 1040	1.83	22
Hachiko: A Dog's Story	1920 × 1036	1.42	17
A Street Cat Named Bob	1920 × 808	1.58	19

In order to ensure the purity of videos, we choose film sources where the subtitles are stored in external.srt files rather than hard-coded into video frames.

Similar to the TVSum dataset [4], we ask annotators to watch the whole video and assess the importance of every shot of the video, so that all video frames in the same shot share the same annotation. Instead of using uniformlength shots in TVSum, we split each video into several shots using the Pearson correlation coefficient [23] between video frames. For different videos, we set different thresholds to ensure that the number of shots does not exceed 100. If the correlation coefficient of two adjacent frames is less than the threshold, the latter frame of the two adjacent frames is treated as a shot switching frame for shot segmentation.

We provide two types of annotations for the dataset, i.e., topic labels and frame-level importance scores. For the topic classification, we provide three classes of labels, namely, animal, character and scenery. For a certain shot, it may be classified into multiple classes at the same time, or none of them. For the video summarization, we provide a coarse-grained score between 0 to 1 for each shot. The important frames in the video that can clearly indicate the video content are marked as 1, and the frames that are not related to the video content are marked as 0.

3.2. Dataset comparison

SumMe [5] and TVSum [4] datasets are used most frequently in existing video summarization literatures. SumMe consists of 25 user videos covering various topics such as holidays and sports, and each video ranges from 1 to 6 minutes. TVSum contains 50 videos covering the topics of news, documentaries, etc. The duration of each video varies from 2 to 10 minutes. It can be seen that the video number of these two datasets is relatively small and only the frame-level importance scores are provided. At the same time, there is less semantic information that can be used in these two datasets and only the short video titles of TVSum can be used for semantic information research sometimes.

In contrast, our dataset, i.e., TopicSum, contains 136 five-minutes-long videos with larger scale, richer data and more diversification. TopicSum not only has frame-level importance score annotations, but also includes topic labels, which can be more in line with practical applications. In addition, TopicSum contains text and audio information, which can support the multimodal task. Table 3 shows the comparison between our dataset and the existing datasets. TopicSum will be published in compliance with regulations at https://github.com/gtz1196/TopicSum.

Table 3
Comparison of our TopicSum dataset and existing datasets.

Dataset	TVSum [4]	SumMe [5]	TopicSum
Video Number	50	25	136
Importance Scores	✓	✓	✓
Topic Labels	-	-	✓
Text Information	✓	-	✓
Audio Information	✓	✓	✓

4. Our method

To address the new task, we propose a multimodal Transformer based method that simultaneously predicts multiple topics and generates the topic-related summaries of a video by fusing multimodal features. Fig. 2 illustrates the overview of our method. It consists of a feature extraction module, a feature learning module, a topic classification module and a frame selection module.

From the logical point of view, the feature extraction module samples the original video and extracts features through three different pre-trained models. The feature learning module adaptively combines features from three different modalities and models the temporal information. The topic classification module classifies the video frames into topic classes. The frame selection module scores the video frames to select relevant frames for generating video summaries.

From the data point of view, the feature extraction module inputs all samples from the same video into a batch and extract their visual features F_v, textual features F_t and audio features F_a using the pre-trained models. The feature learning module first combines a multimodal embedding MLT and the three features F_v, F_t, F_a into a batch as the input of the multimodal Transformer encoder to obtain the fused feature F_m and the updated features of three modalities $\hat{F}_v$, $\hat{F}_t$, $\hat{F}_a$. Then, the feature learning module combines all the fused features from the same video into a batch as the input of the temporal modeling encoder to obtain the updated fused features $\hat{F}_m$. The topic classification module combines all the updated fused features from the same video into a batch as input to predict the topic labels $\hat{Y}_t$, and the frame selection module combines all the updated visual features from the same video into a batch as input to predict the frame-level importance scores $\hat{S}$. Table 4 summarizes the math notations.

4.1. Feature extraction

To utilize rich information in videos for generating high-quality summaries, we integrate the visual, textual and audio information.

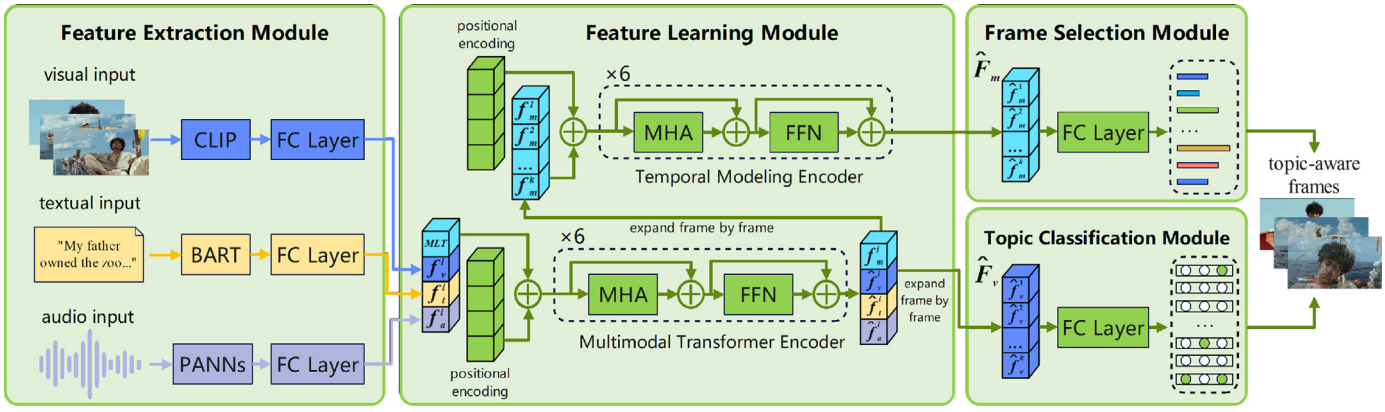

Fig. 2. Overview of our method. During training, the visual, textual and audio features are extracted through three different pre-trained models. Then the feature learning module fuses these three features by a multimodal Transformer encoder, and models the temporal motion in video by a temporal modeling encoder. Finally, the topic classification module predicts the topic labels, and the frame selection module calculates the frame-level importance scores to generate summaries.

Table 4
Math notations.

Notation	Meaning
MLT	Multimodal embedding
F_v	Visual features extracted by CLIP
F_t	Textual features extracted by BART
F_a	Audio features extracted by PANNs
F_m	Fused features generated by the multimodal Transformer encoder
Y_t	Annotated topic labels
S	Annotated importance scores
$\hat{F}_v$	Updated visual features generated by the multimodal Transformer encoder
$\hat{F}_t$	Updated textual features generated by the multimodal Transformer encoder
$\hat{F}_a$	Updated audio features generated by the multimodal Transformer encoder
$\hat{F}_m$	Updated fused features generated by the temporal modeling encoder
$\hat{Y}_t$	Predicted topic labels by the topic classification module
$\hat{S}$	Predicted importance scores by the frame selection module

For each video, we sample at 1fps to obtain k frames and the visual features of the sampled frames are extracted by the Contrastive Language-Image Pre-training model (CLIP) [24], denoted as $F_v = \{f_v^1, \ldots, f_v^k\}$. The subtitle corresponding to each frame is selected from the subtitle file of the video, and encoded by the pre-trained Bidirectional and Auto-Regressive Transformers model (BART) [25], denoted as $F_t = \{f_t^1, \ldots, f_t^k\}$. The audio features are shot-level encoded by the Pre-trained Audio Neural Networks (PANNs) [26], denoted as $F_a = \{f_a^1, \ldots, f_a^k\}$, which means that different video frames in the same shot share the common audio feature. Considering that the dimensions and distributions of these multimodal features are quite different, We append a fully-connected layer following the pre-trained models to fix the dimension of the multimodal features to d_m, and use a layernorm operator to normalize the multimodal features, given by

$$\begin{aligned} f_v^i &= LN(W_v \cdot CLIP(V^i) + b_v), \\ f_t^i &= LN(W_t \cdot BART(T^i) + b_t), \\ f_a^i &= LN(W_a \cdot PANNs(A^i) + b_a), \end{aligned} \tag{1}$$

where V^i represents the i-th video frame, and T^i and A^i represent the subtitle and audio signal of the i-th frame, respectively. In addition to the three features as input to the feature learning module, following the practice in Devlin et al. [27], we take a special multimodal embedding MLT for multimodal fusion, and the output corresponding to this embedding is used as the fused feature for frame selection task.

4.2. Feature learning

4.2.1. Multimodal transformer encoder

We introduce a multimodal Transformer encoder to adaptively fuse the visual feature F_v, the textual feature F_t and the audio feature F_a. We modify the multi-head attention described in [28] to take in inputs from three different modalities. Specifically, for each frame, we combine the features from three modalities into a sequence, and append a multimodal embedding MLT before the multimodal features as the input of the multimodal Transformer encoder. The input sequence is denoted by $X = [MLT, f_v^i, f_t^i, f_a^i]^\top, X \in R^{4\times d_m}$, and the positional encoding is applied to distinguish the features of different modalities. Through the cross-modal attention mechanism, our multimodal Transformer encoder has four outputs: the fused feature, the updated visual feature, the updated textual feature, and the updated audio feature, denoted as $E = [f_m^i, \hat{f}_v^i, \hat{f}_t^i, \hat{f}_a^i]^\top, E \in R^{4\times d_m}$. Similar to [28], the multimodal Transformer encoder is composed of a stack of $N = 6$ identical layers, and the formulas for each layer are as follows:

$$\begin{aligned} Q_h, K_h, V_h &= X \cdot (W_h^Q, W_h^K, W_h^V), h = 1, 2, \ldots, H, \\ SA_h &= softmax\left(\frac{Q_h \cdot K_h^\top}{\sqrt{d_k}}\right) \cdot V_h, h = 1, 2, \ldots, H, \\ MHA &= Concat(SA_1, SA_2, \ldots, SA_H) \cdot W^O, \\ FFN &= max(0, MHA \cdot W_1 + b_1) \cdot W_2 + b_2, \end{aligned} \tag{2}$$

where H is the head number of multi-head attention, and Q_h, K_h, V_h denote the query, key and value matrices of the h-th self-attention layer, respectively. SA_h denotes the result of the h-th self-attention layer, MHA denotes the result of multi-head attention, and FFN denotes the result of position-wise feed-forward networks. $W_h^Q \in R^{d_m\times d_k}$, $W_h^K \in R^{d_m\times d_k}$, $W_h^V \in R^{d_m\times d_k}$, $W^O \in R^{H\cdot d_k\times d_m}$, $W_1 \in R^{d_m\times d_{ff}}$ and $W_2 \in R^{d_{ff}\times d_m}$ denote the learnable parameter matrices, where $d_k = d_m/H$.

4.2.2. Temporal modeling encoder

Capturing the long-range temporal information is of great importance to video understanding, especially to video summarization whose goal is to identify the most representative video frames [9,29]. In our method, an additional temporal modeling encoder is designed to capture the temporal motion in videos. It has the same structure as the multimodal Transformer encoder described in Section 4.2.1, and its input sequence is composed of the fused features, denoted as $F_m = [f_m^1, f_m^2, \ldots, f_m^k]^\top$, $F_m \in R^{k\times d_m}$. The positional encoding is also used to distinguish the features at different times. The output of the temporal modeling encoder is denoted as updated fused features $\hat{F}_m = [\hat{f}_m^1, \ldots, \hat{f}_m^k]^\top, \hat{F}_m \in R^{k\times d_m}$.

4.3. Topic classification

The topic classification module aims to classify the updated visual features $\hat{F}_v = \{\hat{f}_v^1, \ldots, \hat{f}_v^k\}$ into topic classes. Since one video frame may present multiple topics, the topic classification is actually a multi-label classification task. Accordingly, the topic classification module is decomposed into multiple binary classifiers, where each binary classifier predicts whether the input frame belongs to the corresponding topic.

A binary cross-entropy loss is employed for training the classifier for each topic and the total loss function for training the topic classification module is the sum of all the topic-aware losses. Let t denote the type of topic, $Y_t = \{y_t^1, \ldots, y_t^k\}$ denote the ground-truth labels of training video frames belonging to the topic t, and $\hat{Y}_t = \{\hat{y}_t^1, \ldots, \hat{y}_t^k\}$ denote the probabilities of classifying the frames into t. Then the loss function for training the classifier of the topic

t is given by

$$L_t = \frac{1}{k}\sum_{i=1}^{k}(y_t^i \log \hat{y}_t^i + (1-y_t^i)\log(1-\hat{y}_t^i)), \tag{3}$$

where $t \in${animal, character, scenery}. So the total classification loss is expressed as

$$L_{cls} = L_{animal} + L_{character} + L_{scenery}. \tag{4}$$

4.4. Frame selection

The frame selection module takes the updated fused features $\hat{F}_m$ as input and predicts the frame-level importance scores, denoted as $\hat{S} = \{\hat{s}_1, \ldots, \hat{s}_k\}$ where $\hat{s}_i$ indicates the importance score of the i-th frame. The importance scores measure the relevance between the corresponding frames and the original video, where the "relevance" here refers to the correlation between the video content and the high-level semantics. The higher the frame-level importance score is, the more representative the frame is. We provide ground-truth frame-level importance scores in training, denoted by $S = \{s_1, \ldots, s_k\}$. A sparsity constraint loss is employed for limiting the sparsity to force the frame selection module to generate high-quality video summaries. Given the predicted frame-level importance scores $\hat{S}$ and the ground-truth frame-level importance scores S, the sparsity loss is formulated by

$$L_{sum} = \frac{1}{k}\sum_{i=1}^{k}(\hat{s}_i - s_i)^2. \tag{5}$$

4.5. Video summary generation

During testing, given an input video, we first filter out the topic-related video frames via the topic classification module. Then we calculate the shot-level scores by averaging the frame-level importance scores in the same shot. Finally, for each topic, we select the topic-related video shot sets as a video summary.

In the cases where the video frames of a certain topic in the original video are rich, we select the shots by maximizing the total score to ensure that the summary length does not exceed the length of the original video of 15%. This strategy is widely used in the field of video summarization [18]. The maximization step is essentially the 0/1 Knapsack problem, which is known as NP-hard.

In the cases where the video frames of a certain topic in the original video are scarce, i.e. the video frames obtained by topic classification module may be less than 15% of the length of the original video, we directly output the video frames obtained after classification as a topic-aware video summary of this topic. If the original video content does not contain video content of a certain topic, the topic-aware video summary of this topic can not be generated.

4.6. Network training

The topic classification loss L_{cls} and the frame selection loss L_{sum} are combined to train the overall network, given by

$$L = L_{sum} + \alpha L_{cls}, \tag{6}$$

where α is a trade-off parameter. Algorithm 1 summarizes the training procedure using M epochs with N videos. For each iteration, we first extract the multimodal features and fuse them via the multimodal Transformer encoder. Then we obtain the final features via the temporal modeling encoder. Next, we predict the topic labels and the importance scores via the topic classification module and the frame selection module, respectively. Finally, we calculate the training loss to optimize the overall network.

Algorithm 1: Training Process of our model.

Input: N training videos with annotated importance scores and topic labels
Output: Parameters θ of our model
for *epoch* $\in$ *{1, 2, ..., M}* **do**
  for $n \in$ *{1, 2, ..., N}* **do**
    Sample three different modalities at 1fps from the n-th video
    for *every sample in the n-th video* **do**
      Extract visual feature f_v^i by CLIP using Eq. 1
      Extract textual feature f_t^i by BART using Eq. 1
      Extract audio feature f_a^i by PANNs using Eq. 1
      Generate fused feature f_m^i and three updated features $\hat{f}_v^i$, $\hat{f}_t^i$ and $\hat{f}_a^i$ by the multimodal Transformer encoder using Eq. 2
    end
    Generate updated fused features $\hat{F}_m$ by the temporal modeling encoder using the fused features F_m
    Predict topic labels $\hat{Y}_t$ by the topic classification module using the updated visual features $\hat{F}_v$
    Predict importance scores $\hat{S}$ by the frame selection module using the updated fused features $\hat{F}_m$
    Calculate the classification loss L_{cls} using Eq. 4
    Calculate the sparsity loss L_{sum} using Eq. 5
    Calculate the total loss L using Eq. 6
    Optimize the parameters θ of our model: $\theta \leftarrow \theta - \eta \partial L / \partial \theta$
  end
end

5. Experiments

5.1. Experimental setup

5.1.1. Dataset

We conduct experiments on the proposed TopicSum dataset to demonstrate the effectiveness of our method on both quantitative and qualitative evaluations. TopicSum consists of 136 content-rich videos sampled from various movies. The types of movie videos include but are not limited to comedy, family, and biography. Among the 136 videos, we select 85% as the training set and 15% as the testing set, namely 116 videos are used for training and the remaining 20 videos are used for testing. TopicSum provides topic labels, frame-level importance scores and shot split results, where all the video frames in the same shot share the same annotation.

5.1.2. Evaluation metrics

The performance of the topic classification module is evaluated by the accuracy of the classification outputs. Let $t \in \{animal, character, scenery\}$ denote the topic class label. For the i-th video shot, let $\hat{y}_t^i$ denote the prediction result for the topic t, and y_t^i denote the ground-truth label. The classification accuracy of the topic t for each video is calculated by

$$Acc_t = \frac{shot\ number\ of\ (y_t^i = \hat{y}_t^i)}{shot\ number\ of\ the\ video}, \tag{7}$$

The average classification accuracy of the three topics is also reported for evaluation.

We follow the protocol in [9] to evaluate the generated topic-aware video summaries, namely measuring the agreement between the generated summaries and the ground-truth summaries using F-score. For the topic t, let X_t denote the generated summary for a video and Y_t denote the ground-truth summary. For

Table 5
Comparison with the previous state-of-the-art methods, where *P*, *R* and *F* indicate the precision, recall and F-score of summaries, respectively.

Method	*P*(%)	*R*(%)	*F*(%)
DR-DSN [30]	39.20	45.70	40.00
SUM-FCN [11]	41.50	46.50	41.80
VASNet [29]	43.78	**49.72**	44.34
$\text{DSNet}_{anchor\_based}$ [12]	50.37	43.17	44.28
$\text{DSNet}_{anchor\_free}$ [12]	50.67	43.67	44.70
Ours	**51.65**	45.04	**45.89**

each video, the F-score of the topic t is denoted as F_t, calculated by

$$P_t = \frac{overlapped\ duration\ of\ X_t\ and\ Y_t}{duration\ of\ X_t},$$
$$R_t = \frac{overlapped\ duration\ of\ X_t\ and\ Y_t}{duration\ of\ Y_t}, \quad (8)$$
$$F_t = \frac{2 \cdot P_t \cdot R_t}{P_t + R_t},$$

where P_t and R_t represent the precision and recall of the topic t for each video, respectively. We also report the average precision, recall and F-score of the three topics.

5.1.3. Implementation details

The training and testing process is implemented using PyTorch. The dimension of the multimodal features d_m is set to 512, the head number H of the multi-head attention is set to 8, and the dimension of the position-wise feed-forward networks d_{ff} is set to 2048. The value of the trade-off parameter α is set to 1. We use the Adam optimizer and the learning rate is set to 0.0001.

5.2. Comparison with state-of-the-art methods

Topic-aware video summarization task is newly proposed, and the existing methods and datasets of video summarization are no longer tailored to this new task. Therefore, when we compare our method with other state-of-the-art methods on our TopicSum dataset, we ignore the topic labels and directly generate video summaries based on the importance scores. The comparison results between our method and the other state-of-the-art methods are shown in Table 5. We observe that our method achieves the highest F-score of 45.89%, which demonstrates that our method makes better use of the information in the video and generates high-quality video summaries by using multimodal features.

Table 6
Quantitative evaluation results of topic classification and video summary generation. *Acc* indicates the topic classification accuracy. *P*, *R* and *F* indicate the precision, recall and F-score of summaries, respectively.

Topic	*Acc*(%)	*P*(%)	*R*(%)	*F*(%)
Animal	86.99	68.18	61.18	63.24
Character	86.22	63.46	64.84	64.11
Scenery	91.27	51.21	40.78	41.05
Average	88.16	60.95	55.60	56.14

5.3. Quantitative evaluation

5.3.1. Quantitative results

We show the evaluation results of video summaries on different topics in Table 6, including the accuracy of topic classification *Acc*, the precision rate of video summaries *P*, the recall rate of video summaries *R* and the F-score *F*.

From the results shown in Table 6, we observe that the overall accuracy of topic classification reaches 88.16%, and the lowest topic classification accuracy of the three topics is 86.22%. This indicates that our method is capable of accurately classifying most video segments. It is also interesting to observe that the scenery topic achieves the best performs, probably due to that video frames of the scenery topic are establishing shots, such as mountains (e.g. some shots of "The Chronicles of Narnia") and sea (e.g. some shots of "Life of Pi"), with more distinctive features for classification.

In terms of video summary generation, it can be seen from the metrics of precision *P*, recall *R* and F-score *F* that our method succeeds in generating representative video summaries on multiple topics. Compared with the animal and character summaries, the scenery summaries perform worse on the F-score *F*. The possible reason is that the information in the scenery frames is relatively less, and its importance to the whole video is difficult to be evaluated, leading to the lower precision and recall, even if its topic classification performance is well.

5.3.2. Ablation study of input modalities

To evaluate the effect of multimodal features, we separately evaluate the performance of our method when removing the visual feature ("w/o video"), the textual feature ("w/o text") and the audio feature ("w/o audio"). Table 7 shows the results of the ablation studies, and we can have the following observations. First, the method using three modalities has better overall performance than the method removing one modality, which verifies that all these three modalities are beneficial to topic-aware video summarizaiton. Second, the results of removing visual feature are signif-

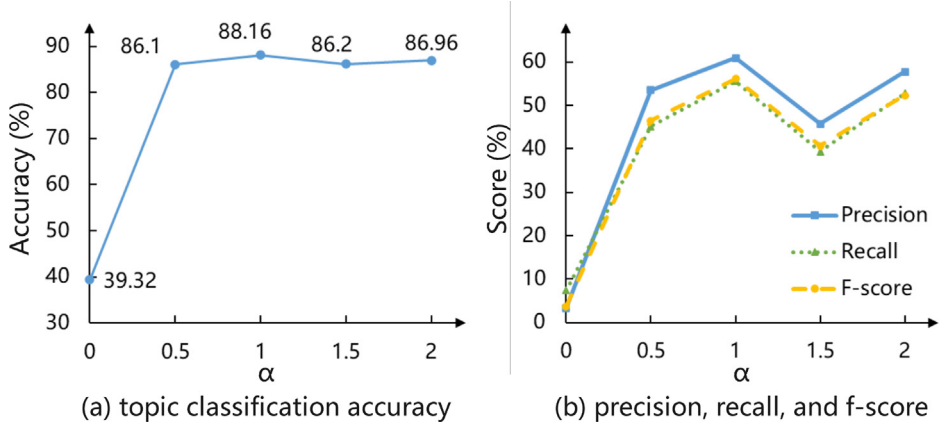

Fig. 3. Results of experiment with different trade-off parameter weight α. The horizontal axis represents different values of α, and the vertical axis represents the topic classification accuracy *Acc*, the precision *P*, the recall *R*, and the F-score *F*.

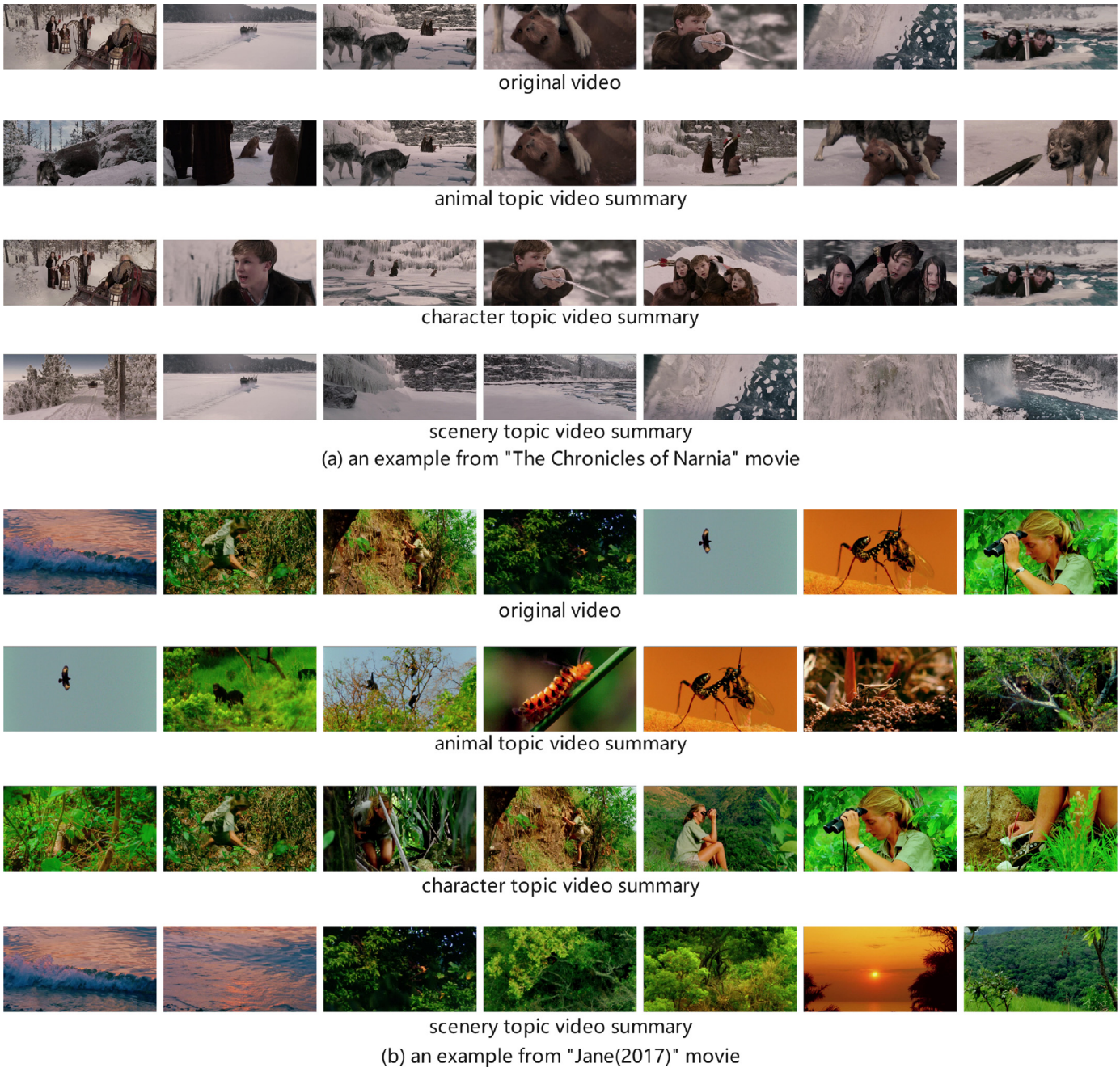

Fig. 4. Two examples of multiple topic-aware video summaries in our TopicSum dataset.

Table 7
Results of ablation studies using different modalities. *Acc* indicates the topic classification accuracy. *P*, *R* and *F* indicate the precision, recall and F-score of summaries, respectively.

Method	*Acc*(%)	*P*(%)	*R*(%)	*F*(%)
w/o video	71.05	15.64	13.80	14.08
w/o text	88.06	55.01	51.40	51.25
w/o audio	84.73	52.38	52.48	50.46
Ours	**88.16**	**60.95**	**55.60**	**56.14**

icantly lower than those of removing other features, showing that the visual information plays a leading role in video summarization. Finally, video subtitles and audio provide supplemental semantic information that helps improving the performance.

5.3.3. Sensitivity analysis of parameter

We also evaluate the effect of the trade-off parameter α on the performance of video summarization, and the results are shown in Fig. 3. The horizontal axis represents different values of α, and the vertical axis represents *Acc*, *P*, *R* and *F*. When α is equal to 0, the performance is worst. As the value of α increases to 1, the classification performance and the summarization performance are both improved, which demonstrates the importance of the classification loss to topic-aware video summarization.

5.4. Qualitative evaluation

5.4.1. Qualitative results

We show two examples of multiple topic-aware video summaries from “The Chronicles of Narnia” movie and “Jane(2017)”

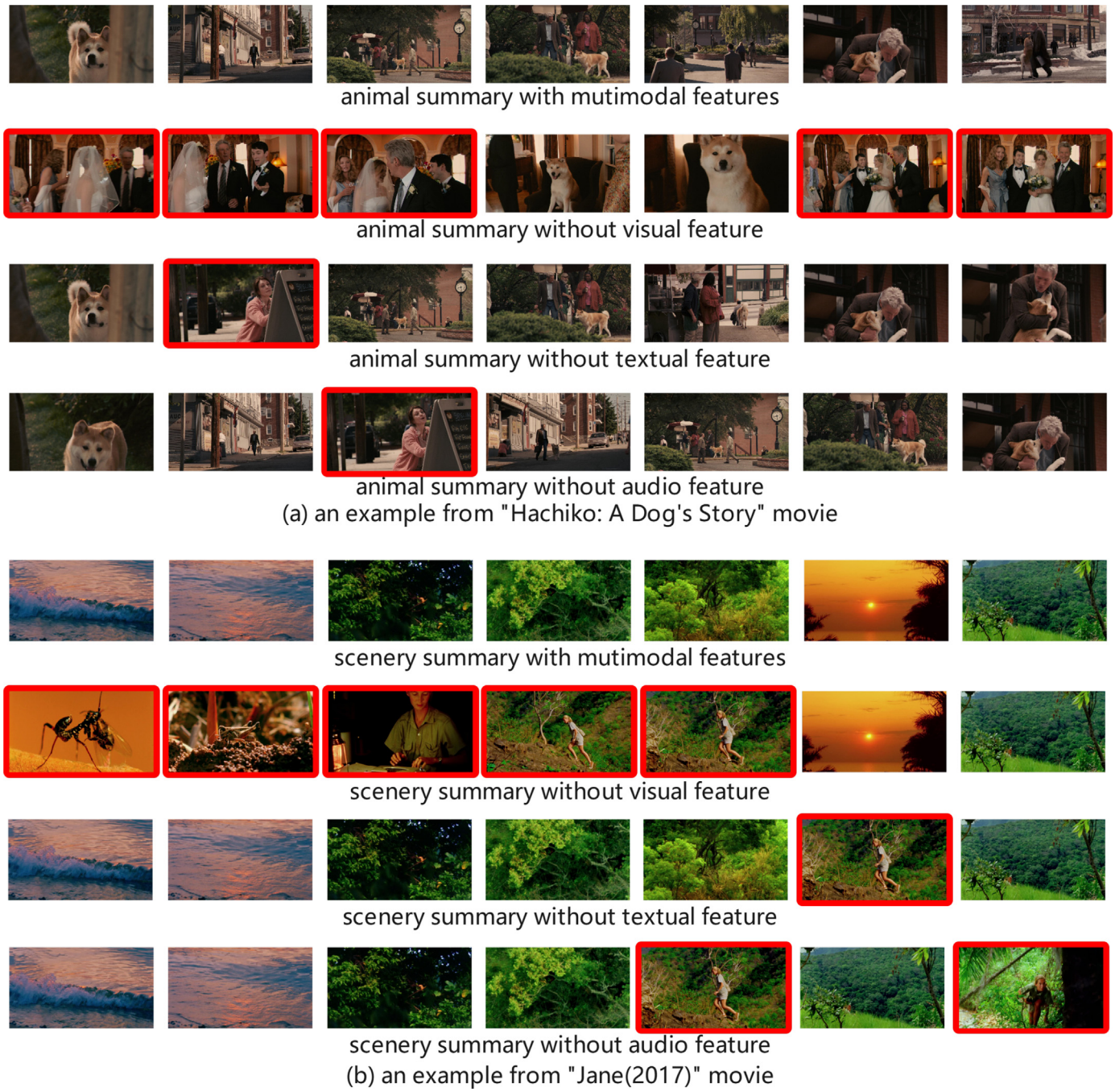

Fig. 5. Qualitative evaluation results of ablation study. The example videos are from "Hachiko: A Dog's Story" movie and "Jane(2017)" movie, which respectively select animal topic and scenery topic. Red boxes indicate video frames not related to the topic in the summary.

movie in Fig. 4. For the first example shown in Fig. 4(a). The original video tells about three children crossing a frozen waterfall while being hunted by wolves. During the battle with wolves, the waterfall melts and three children accidentally fall into the water. From the content of the original video, we know that the main story most related to animal is that the battle between wolves and three children, and the summary about animal covers the animal-related content in the original video. The summary about character focuses on the scenes of how the three children pass through the frozen waterfall and finally fall into the water, which also tells the main story of the video from a character-related perspective. For the summary about scenery, there are some beautiful or vast shots in the story, and contains the whole process of the waterfall melting. Fig. 4(b) shows the second example. The original video tells about the animals and scenery Jane saw when she first came to Africa to observe chimpanzees in the primeval forest, and the three topic-aware video summaries tell about the rich animals in the primeval forest, Jane's process of observing chimpanzees, and the beautiful scenery along the way, respectively.

5.4.2. Qualitative evaluation of ablation study

In order to analyze our ablation experiments more intuitively, we use two examples from "Hachiko: A Dog's Story" movie and "Jane(2017)" movie to show its different video summaries when different features are used as input. For the first example shown in Fig. 5(a). The original video contains scenes of dog Hachiko accom-

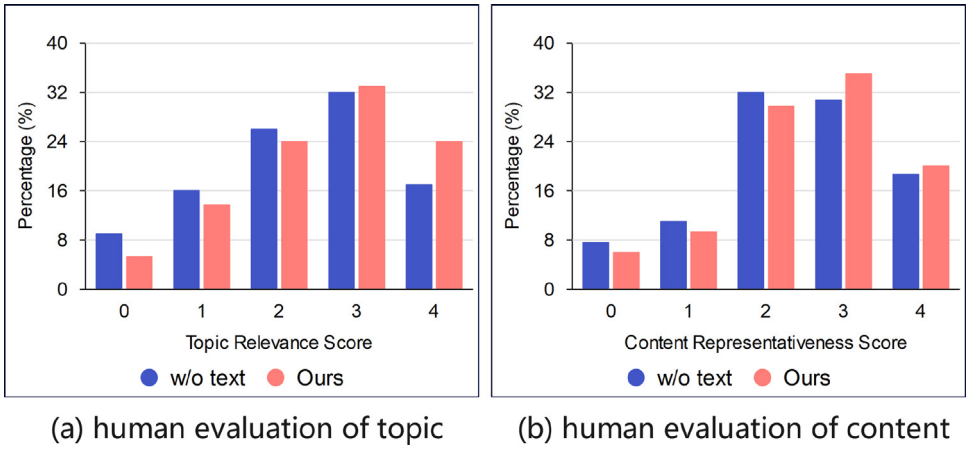

Fig. 6. Human evaluation results of ablation study. The topic relevance ranges from 0 (not relevant to the topic) to 4 (perfectly relevant to the topic). The content representativeness ranges from 0 (not representative) to 4 (very representative). The horizontal axis indicates the score and the vertical axis represents the percentage of each score.

panying its owner to and from work. It can be seen that the summary generated without visual feature contains a large number of frames unrelated to animal topic (shown in red boxes), and does not cover the main story of the video. Although the summaries generated without textual feature or audio feature are better than that without visual feature, there still exists the misclassification of animal topic. The quality of the summary generated by utilizing multimodal features is the best, where not only the topic classification is correct, but also the generated summary covers the main story that the original video wants to express. The second example shown in Fig. 5(b) shows similar results. The summary generated by utilizing multimodal features successfully shows the beautiful scenery Jane saw while working in the primeval forest in Africa, and the lack of modal leads to the misclassification of scenery topic and reduces the quality of the summary.

5.5. Human evaluation

To further evaluate the quality of video summaries, we perform human evaluation by recruiting 15 users from different backgrounds to conduct quality assessment.

5.5.1. Analysis of multimodal features

Since the performance of "w/o text" is better than that of "w/o video" and "w/o audio", we compare the human evaluation results between the summaries generated by "w/o text" and our method. For each test video, we show the original video, the summary generated by our method and the summary generated by "w/o text" method, the users are asked to assess the summary in terms of the topic relevance and the content representativeness. The topic relevance reflects the extent to which the summary is relevant to the topic and ranges from 0 (not relevant to the topic) to 4 (perfectly relevant to the topic). The content representativeness measures how the summary represents the content of the original video and also ranges from 0 (not representative) to 4 (very representative).

Fig. 6 shows the results of human evaluation where the horizontal axis indicates the score of topic relevance (a) or the score of content representativeness (b) and the vertical axis represents the percentage of each score. It is interesting to observe that most summaries generated by our method are given high scores (i.e., 3 and 4) on both topic relevance and content representativeness, which shows the superiority of our method on generating high-quality summaries on multiple topics. Compared with "w/o text", fewer video summaries of "Ours" are evaluated low scores (i.e., 0 and 1) and more video summaries of "Ours" are evaluated high scores (i.e., 3 and 4), which further validates the merit of integrating textual feature for summarization.

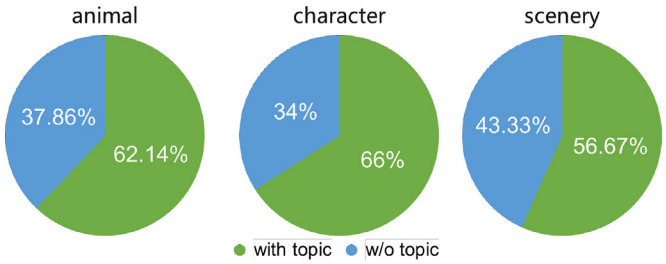

Fig. 7. Human evaluation results of topic awareness. For each topic, "with topic" represents the percentage of summaries with topic information that are chosen by users and "w/o topic" represents the percentage of summaries without topic information that are chosen by users.

5.5.2. Analysis of topic awareness

We also conduct human evaluation to analyze the topic awareness of videos summaries generated by our method. That is to say, we evaluate whether the topic-aware video summaries are preferred by users. Concretely, we investigate in advance the interested topic for each user, and then show each user the video summary on his/her interested topic and the video summary without topic information. The users are asked to choose the video summaries they are more interested in.

Fig. 7 shows the human evaluation results of topic awareness. For each topic, "with topic" represents the percentage of summaries with topic information that are chosen by users and "w/o topic" represents the percentage of summaries without topic information that are chosen by users. It is obvious that the topic-aware video summaries are more preferred by users for all the topics. It also be seen that compared with the topics of animal and scenery, the character topic-related video summaries attract more attention from users. The possible reason is that our dataset is collected from movies, and the shots containing characters are often related to the main story, so users prefer video summaries on the character topic.

6. Conclusion

We have presented a novel topic-aware video summarization task, and a multimodal Transformer based model as the baseline method for this new task. Additionally, we have built a benchmark

dataset, called TopicSum, which is collected and annotated with both topic labels and frame-level importance scores. The topic-aware video summarization can generate multiple topic-related synopsis to represent video content from various perspectives and satisfy the interests of users compared with conventional video summarization. However, the topics in TopicSum are relatively limited compared with user interests.

In the future work, we are going to expand the TopicSum dataset by including more videos and annotations with diversified topics, with a view to increasing attention to the impact of user interests on video summarization. Moreover, we will build a bridge between our video topics and the existing video tags used by various video platforms to help users intuitively understand different topics and further improve the practicality of our method in real-world scenarios.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported in part by the Natural Science Foundation of China (NSFC) under Grant No 62072041.

References

[1] B. Zhao, E.P. Xing, Quasi real-time summarization for consumer videos, CVPR, 2014.
[2] W. Zhu, J. Lu, Y. Han, J. Zhou, Learning multiscale hierarchical attention for video summarization, Pattern Recognit 122 (2022) 108312.
[3] G. Liang, Y. Lv, S. Li, S. Zhang, Y. Zhang, Video summarization with a convolutional attentive adversarial network, Pattern Recognit 131 (2022) 108840.
[4] Y. Song, J. Vallmitjana, A. Stent, A. Jaimes, Tvsum: Summarizing web videos using titles, CVPR, 2015.
[5] B. Gong, W.-L. Chao, K. Grauman, F. Sha, Diverse sequential subset selection for supervised video summarization, Adv Neural Inf Process Syst 27 (2014) 2069–2077.
[6] S.E.F. de Avila, A.P.B. Lopes, A. da Luz Jr, A. de Albuquerque Araújo, VSUMM: a mechanism designed to produce static video summaries and a novel evaluation method, Pattern Recognit Lett 32 (1) (2011) 56–68.
[7] R. Laganière, R. Bacco, A. Hocevar, P. Lambert, G. Païs, B.E. Ionescu, Video summarization from spatio-temporal features, in: Proceedings of the 2nd ACM TRECVid Video Summarization Workshop, 2008, pp. 144–148.
[8] Z. Ren, J. Yan, B. Ni, B. Liu, X. Yang, H. Zha, Unsupervised deep learning for optical flow estimation, AAAI, 2017.
[9] K. Zhang, W.-L. Chao, F. Sha, K. Grauman, Video summarization with long short-term memory, ECCV, 2016.
[10] B. Zhao, X. Li, X. Lu, Hsa-rnn: Hierarchical structure-adaptive RNN for video summarization, CVPR, 2018.
[11] M. Rochan, L. Ye, Y. Wang, Video summarization using fully convolutional sequence networks, ECCV, 2018.
[12] W. Zhu, J. Lu, J. Li, J. Zhou, Dsnet: a flexible detect-to-summarize network for video summarization, IEEE Trans. Image Process. 30 (2020) 948–962.
[13] J.A. Ghauri, S. Hakimov, R. Ewerth, Supervised video summarization via multiple feature sets with parallel attention, in: 2021 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2021, pp. 1–6s.
[14] E. Apostolidis, E. Adamantidou, A.I. Metsai, V. Mezaris, I. Patras, AC-SUM-GAN: connecting actor-critic and generative adversarial networks for unsupervised video summarization, IEEE Trans. Circuits Syst. Video Technol. 31 (8) (2020) 3278–3292.
[15] T. Liu, Q. Meng, J.-J. Huang, A. Vlontzos, D. Rueckert, B. Kainz, Video summarization through reinforcement learning with a 3D spatio-temporal u-net, IEEE Trans. Image Process. 31 (2022) 1573–1586.
[16] S.-H. Zhong, J. Lin, J. Lu, A. Fares, T. Ren, Deep semantic and attentive network for unsupervised video summarization, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 18 (2) (2022) 1–21.
[17] A. Sharghi, A. Borji, C. Li, T. Yang, B. Gong, Improving sequential determinantal point processes for supervised video summarization, ECCV, 2018.
[18] H. Wei, B. Ni, Y. Yan, H. Yu, X. Yang, C. Yao, Video summarization via semantic attended networks, AAAI, 2018.
[19] S. Xiao, Z. Zhao, Z. Zhang, X. Yan, M. Yang, Convolutional hierarchical attention network for query-focused video summarization, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 12426–12433.
[20] M. Narasimhan, A. Rohrbach, T. Darrell, Clip-it! language-guided video summarization, Adv Neural Inf Process Syst 34 (2021).
[21] Y. Yuan, T. Mei, P. Cui, W. Zhu, Video summarization by learning deep side semantic embedding, IEEE Trans. Circuits Syst. Video Technol. 29 (1) (2017) 226–237.
[22] H. Li, Q. Ke, M. Gong, T. Drummond, Progressive video summarization via multimodal self-supervised learning, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 5584–5593.
[23] J. Benesty, J. Chen, Y. Huang, I. Cohen, Pearson correlation coefficient, in: Noise reduction in speech processing, Springer, 2009, pp. 1–4.
[24] 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: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.
[25] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, arXiv preprint arXiv:1910.13461 (2019).
[26] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, M.D. Plumbley, PANNs: large-scale pretrained audio neural networks for audio pattern recognition, IEEE/ACM Trans Audio Speech Lang Process 28 (2020) 2880–2894.
[27] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, NIPS, 2017.
[29] J. Fajtl, H.S. Sokeh, V. Argyriou, D. Monekosso, P. Remagnino, Summarizing videos with attention, in: Asian Conference on Computer Vision, Springer, 2018, pp. 39–54.
[30] K. Zhou, Y. Qiao, T. Xiang, Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward, AAAI, 2018.

Yubo Zhu received the B.S. degree in computer science from Beijing Institute of Technology, Beijing, China, in 2021. He is currently working toward the M.S. degree in computer science with the School of Computer Science and Technology, Beijing Institute of Technology. His research interests include computer vision and video summarization.

Wentian Zhao received the B.S. degree in computer science from the Beijing Institute of Technology, Beijing, in 2017, where he is currently pursuing the Ph.D. degree with the Beijing Laboratory of Intelligent Information Technology, School of Computer Science and Technology. His research interests include computer vision and natural language processing.

Rui Hua received the B.S. degree in digital media technology from Beijing Institute of Technology, Beijing, China, in 2019, and the M.S. degree in software engineering from Beijing Institute of Technology, Beijing, China, in 2021. Her research interests include computer vision and video summarization.

Xinxiao Wu received the B.S. degree in computer science from the Nanjing University of Information Science and Technology, Nanjing, China, in 2005, and the Ph.D. degree in computer science from the Beijing Institute of Technology (BIT), Beijing, China, in 2010. From 2010 to 2011, she was a Postdoctoral Research Fellow with Nanyang Technological University, Singapore. She is currently a Full Professor with the School of Computer Science, BIT. Her research interests include machine learning, computer vision, and video analysis and understanding. She has served on the Editorial Boards for the IEEE Transactions on Multimedia.