

SPECIFICATION

MULTIMODAL HETEROGENEOUS FEATURE FUSION-BASED COMPACT VIDEO EVENT DESCRIPTION METHOD

TECHNICAL FIELD

The present invention relates to the field of computer vision, to the field of video description, and in particular, to a multimodal heterogeneous feature fusion-based compact video event description method.

BACKGROUND

In the aspect of application scope and scale, computer vision is the most extensive and common in the application of artificial intelligence, has already penetrated into various aspects of daily life and work, involving network security, system evaluation, monitoring, intelligent machines, etc., plays an important role in the development of human civilization, and promotes the development and progress of society. In image identification of computer vision and artificial intelligence, there are mainly static images and dynamic images. The static images mainly include pictures, etc., while the dynamic images mainly include videos, etc. For a video description task, it is an important research field at present to identify the relationship between these events in the video and describe all the events, especially in the specific applications such as video monitoring, network security and system evaluation.

At present, due to the high development of Internet and terminal applications, involved videos in an open domain, long videos and dense events are often numerous, and are constantly updated. Because long videos in open domains usually include multiple events, which involve multiple behaviors and entities, a reasonable description way should be to locate and describe each event. A general video description task is to generate only one text description for a video, which is obviously insufficient when the video contains multiple events or complex scenes.

In addition, due to the fact that the image description task generates a text description for each picture, if applying the image description method to a video is to describe a specific video frame, and if each frame is described, a high calculational cost is required. Moreover, because the generated descriptions are also too dense, post-processing is also needed to integrate the descriptions, which increases a lot of workload..

SPECIFICATION

In the field of image or video description, the traditional machine learning methods can no longer meet the requirements, and now Convolutional Neural Networks (CNN) are commonly used. One of the biggest reason is that the feature design of image data, i.e., feature description, has always been the difficult point in researches of computer vision.

In feature description, the focus is on feature fusion. Feature-level fusion is the fusion of abstract features, which may be divided into early fusion and late fusion. In early fusion, features are fused first and then sent to a model for prediction, while in late fusion, features are sent to the model for prediction first and then fused for uniform scoring. Traditional feature fusion algorithms are mostly early fusion algorithms, which generally involves summing or cascading multi-modal features. However, these simple linear fusion methods can't have the complex feature interaction capability between multi-modal features. At present, too many parameters and insufficient interaction of multi-modal features during fusion are the main problems in the field of multi-modal fusion. Moreover, most of the traditional fusion algorithms fail to meet the requirements of multi-scene applications in reality, and lack the generalization capability of models.

Although there have been many researches related to video description, at present, there are still many challenges for video description. Not only because the video description task itself demands high encoding and decoding capability of a model, but also because some current objective factors limit the development of video description, the specific reasons may be summarized as follows: (1) at present, the demand for video description methods of videos in open domains, long videos and videos of dense events is increasing day by day, and at present, a lot of researches focus on image description or short video description only containing single events; (2) the current video description methods rely too much on the intuitive understanding of known data and generate simple co-occurrence descriptions, and lack the reasoning process for potential information, which leads to a limitation on the understanding capability of the model; and (3) most video description methods haven't fully utilized the existing data, and only utilize single modal data for analysis, limiting the structure and performance of the model.

SUMMARY

In order to effectively solve the defect that meaningless descriptions, such as co-occurrence descriptions, are frequently generated in video description methods of long videos and dense events, and meet the requirements of multi-scene applications at the same time, the present invention provides a multimodal heterogeneous feature fusion-based compact video event

SPECIFICATION

description method.

The present invention claims following technical schemes:

The present invention provides a multimodal heterogeneous feature fusion-based compact video event description method that includes following steps of:

feature extracting: selecting a long video in an open domain for clipping to obtain a corresponding video frame by utilizing an I3D convolutional network, and extracting a dynamic visual feature F_m in the video frame to acquire a dynamic visual feature vector; and extracting an audio prosody feature vector in a video segment by utilizing a VGGish model;

scene graph feature embedding: semantically representing scene object information in the video frame, generating a scene graph, obtaining an entity code, an attribute code and a relationship code, and performing graph embedding on corresponding feature vectors in the entity code, the attribute code and the relationship code through graph convolution to obtain a scene feature embedding vector;

triple multi-modal loop fusing: taking the scene graph feature embedding vector, the dynamic visual feature vector, and the audio feature vector as input and performing triple multi-modal loop fusion on these vectors, then normalizing the output vector along the channel using the normalized function, and using Dropout to regularize and the regularized output vector is input into a decoder to prepare for reconstructing decoder;

describing reconstruction decoder, including a pre-trained language dictionary decoder and a description reconstruction algorithm, detecting a video dense event by means of the description reconstruction algorithm; decoding a description of the generated video scene event by the pre-trained language dictionary decoder, where the pre-trained language dictionary decoder refers to a module that utilizes a pre-trained language dictionary to decode the description of the video scene event.

Furthermore, for the dynamic visual feature F_m , $F_m = [f_1^m, f_2^m, \dots, f_N^m]$, where N is the number of extracted video clips, and the whole corresponding dimension is $T_v \times d_v$; the dimension corresponding to the audio prosody feature vector is $T_a \times d_a$, and the audio feature vector is converted into a 128-dimensional feature vector with semantics.

Furthermore, in the process of scene graph feature embedding, the video frame is an input, firstly, object position information about object bounding box coordinates, etc. in a video frame picture is detected by means of the Mask R-CNN, a position of an object is determined to obtain an object region; then an original picture of the video frame and the object region are respectively

SPECIFICATION

sent to a visual scene graph generator to generate a scene graph, visual information such as information about the video frame picture and information about the detected object position are semantically represented to obtain an entity code, an attribute code and a relationship code, these codes are then sent to a joint embedding layer for joint representation to obtain semantic information, the semantic information is embedded into corresponding feature vectors in the entity code, the attribute code and the relationship code; later graph embedding is performed on the feature vectors by means of graph convolution to obtain a scene graph feature embedding vector; and the association between entities is reinforced inferenced in this process.

Furthermore, after the triple multi-modal loop fusing, the present invention further includes:

adaptive multi-modal data balancing: extracting features of the dynamic visual feature vector and the audio prosody feature vector in a time sequence dimension by utilizing a fully convolutional network, and using specific values to predict time sequence boundaries and confidence scores of events; secondly, sending the predicted events to a common pool, and ranking the events in a descending order and selecting events with high confidence scores; and clustering the events with high confidence scores, and selecting various corresponding clustering centers to form a final event set.

Furthermore, the specific values refer to a center position $center = p + \sigma(c)$, where $\sigma(\cdot)$ is a *sigmoid* function configured to scale a value of c to a range of $[0, 1]$ to determine a right neighborhood of a center point p ; an event duration $length = anchor \cdot e$, where e is a scaling coefficient; and a confidence score $confidence = \sigma(o)$.

Furthermore, the fully convolutional network adopts one-dimensional convolution to extract features in the time sequence dimension, the size of a convolution kernel of the first layer is k , the sizes of both of convolution kernels of the second layer and the third layer are 1, the activation function of each of the first two layers is Relu, and the activation function of the last layer is *sigmoid*.

Furthermore, in the process of triple multi-modal loop fusing, the normalized function refers to a Softmax function.

In the process of triple multi-modal loop fusion, dimensions of $T_v \times d_v$ and the dynamic visual feature vector and the audio feature vector of $T_v \times d_v$ are first sent to a self-attention layer for feature representation.

Furthermore, events with the top 100 high confidence scores are selected, and the K-Means algorithm based on Euclidean distance is adopted as a clustering algorithm.

SPECIFICATION

Furthermore, a pre-training method for the language dictionary includes, converting annotated description text into a scene graph, then performing scene graph feature embedding by means of a graph convolutional network, calculating the obtained vector being computed with the pre-trained language dictionary to re-code the scene graph feature, then sending the re-coded vector to a decoder for reconstruction and then for text description, and supervising model learning by comparing the difference between the reconstructed text description and the annotated description text, the annotated description text being data for pre-training stored in a data set;

the pre-trained language dictionary refers to that the pre-trained language dictionary D is a matrix with a dimension of $d \times L$;

the re-coding refers to the calculation of the re-coding $\varphi(\cdot)$, and the process is as follows:

$$\hat{x} = \varphi(x, D) = \sum_{l=1}^L \text{softmax}(D^T x_l) d_l$$

where d_l represents the vector in the column 1 of D , and D herein has the function of encoding and embedding human language habits.

Furthermore, in the description reconstruction decoder step, the description reconstruction algorithm specifically includes:

Description Reconstruction Algorithm

Input: start time P and end time Q of a video; event set E , each event containing attributes: central time point m of the event, event duration l , confidence score s , and description z corresponding to the event

Output: candidate description set Z

1. According to m and l , calculating start time s and end time e of each event:

$$s = m - \frac{l}{2}, e = m + \frac{l}{2};$$

2. Calculating tIOU between events, and de-duplicating overlapping events;

3. Ranking the events in an ascending order of start time and selecting an event with the earliest start time as a reference event k ;

4. Selecting events with start time within $[k_s, k_s + \mathcal{E}]$ from a remaining event set $E - k$ to form a start event set E_1 together;

SPECIFICATION

5. Selecting events with start time within $[E_{1_e} - \mathcal{E}, E_{1_e} + \mathcal{E}]$ from a remaining event set $E - E_1$ to form an event sequence set $\hat{E}$ together with corresponding events in E_1 ;
 6. Repeating step 5 and updating $\hat{E}$ until the end time of the selected events is within the range of $[Q - \mathcal{E}, Q]$, updating $\hat{E}$, and ending selection;
 7. Merging text descriptions corresponding to every event sequence in $\hat{E}$ and adding the merged text description to a candidate description set Z .
-

The present invention further provides a fusion method considering both parameter quantity and interaction intensity that the fusion method is configured to realize the triple multi-modal loop fusion described above, and enable multiple features to be deeply fused on the premise of not generating high-dimensional features.

The present invention further provides a video dense event description device based on multi-modal heterogeneous feature fusion that includes:

a feature extraction module configured to select a long video in an open domain for clipping to obtain a corresponding video frame by utilizing an I3D convolutional network, and extract a dynamic visual feature F_m in the video frame to acquire a dynamic visual feature vector; and extract an audio prosody feature vector in a video segment by utilizing a VGGish model;

a scene graph feature embedding module configured to semantically represent scene object information in the video frame, generate a scene graph, obtain an entity code, an attribute code and a relationship code, and perform graph embedding on corresponding feature vectors in the entity code, the attribute code and the relationship code through graph convolution to obtain a scene feature embedding vector;

a triple multi-modal loop fusion module configured to take the scene graph feature embedding vector, the dynamic visual feature vector and the audio feature vector as input and perform triple multi-modal loop fusion on these vectors, normalize the output vector along the channel using the normalized function, and use Dropout to regularize and the regularized output vector is input into a decoder to prepare for reconstructing decoder, where the triple multi-modal loop fusion module includes a pre-established fusion method considering both parameter quantity and interaction intensity, which enables multiple features to be deeply fused on the premise of not generating high-dimensional features; and

a description reconstruction decoder module configured to detect a video dense event by means of a description reconstruction algorithm; and decode a description of the generated video scene event by a pre-trained language dictionary decoder, where the pre-trained language

SPECIFICATION

dictionary decoder refers to a module that utilizes a pre-trained language dictionary to decode the description of the video scene event.

Furthermore, the video dense event description device further includes an adaptive multi-modal data balancing module configured to extract features of the dynamic visual feature vector and the audio prosody feature vector in a time sequence dimension by utilizing a fully convolutional network, and predict time sequence boundaries and confidence scores of events with specific values; secondly, send the predicted events to a common pool, and rank the events in a descending order and select events with high confidence scores; and finally, cluster the events with high confidence scores, and select various corresponding clustering centers to form a final event set.

Compared with the existing technology, the present invention has the following advantages:

For description of the video dense event, features are extracted by utilizing the I3D convolution network and the VGGish model, the video description task is refined, so that the problem of locating each event in a long video description is solved. The way using the scene graph feature embedding is proposed for the first time in the video dense event description method, and multi-modal loop fusion is realized by the triple multi-modal loop fusion module designed by the present invention, so that the video dense description task describes the correlation between event objects, the effect is better, and the defect that a lot of current researches focus on image description or short video description only containing single event is solved.

(2) The present invention is equipped with adaptive multi-modal data balancing, so the dynamic vision feature and audio prosody feature are fitted with each other to ensure the rationality of event extraction and improve the generalization capability of a model.

(3) The present invention also reconstructs a decoder, designs a description reconstruction algorithm to detect video events, and generates descriptions of video scene events according to a pre-trained language dictionary, effectively solving the defect that meaningless descriptions, such as co-occurrence descriptions and others, are frequently generated in video description methods, and effectively utilizing multi-modal information to mine the relationship between scene events.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a schematic step diagram of a multimodal heterogeneous feature fusion-based compact video event description method provided by the present invention.

Fig. 2 is a configuration program diagram of a video dense event description device based on

SPECIFICATION

multi-modal heterogeneous feature fusion provided by the present invention.

Fig. 3 is a configuration program diagram of another video dense event description device based on multi-modal heterogeneous feature fusion provided by the present invention.

Fig. 4 is a schematic step diagram of another multimodal heterogeneous feature fusion-based compact video event description method provided by the present invention.

Fig. 5 is a flowchart of another multimodal heterogeneous feature fusion-based compact video event description method provided by the present invention.

Fig. 6 is a schematic diagram showing scene graph embedding and generation of another multimodal heterogeneous feature fusion-based compact video event description method provided by the present invention.

Fig. 7 is a schematic diagram of a triple multi-modal loop fusion module of another multimodal heterogeneous feature fusion-based compact video event description method provided by the present invention.

Fig. 8 is a schematic diagram showing language dictionary training of another multimodal heterogeneous feature fusion-based compact video event description method provided by the present invention.

DETAILED DESCRIPTION

To make those having ordinary skill in the art better understand the schemes of the present invention, the technical schemes in the embodiments of the present invention will be described clearly and completely below in combination with the drawings in the embodiments of the present invention. Obviously, the described embodiments are only a part of the embodiments of the present invention, but not all of the embodiments. Based on the embodiments of the present invention, all other embodiments obtained by those having ordinary skill in the art without contributing creative labor should fall within the protection scope of the present invention.

In order to facilitate the understanding of the technical schemes of the present invention, the technical terms involved in the present invention will be explained in detail below.

The video dense event description task is a further refinement of a video description task, which just solves the problem of locating each event in a long video description. The video dense event description task is to first locate video segments containing key events in a video, including timestamps when the events start and end, then describe the obtained segments, and finally integrate multi-sentence descriptions. This task can ensure the correctness of the generated

SPECIFICATION

descriptions in time sequence, meet the requirements of multi-event descriptions, and is more challenging than the general video description task.

Scene graph is of a data structure of a directed graph, which may be defined as a tuple (triplet) $G=(O, E)$, where O is a group of objects detected in an image, represented as nodes in the graph. Each object has a form of $O_i = (C_i, A_i)$, where C_i and A_i are category and attribute of the object respectively. E is a group of directed edges, indicating paired relationships between objects, which may be verbs, spaces, prepositions and comparative degrees. The scene graph may contain structured semantic information about the image, including knowledge of current objects, object attributes and pairwise relationships. Therefore, the scene graph may provide useful prior knowledge for image retrieval, video question answering, image generation and other visual tasks.

Graph convolution is a calculation method in the structure of a graph convolution neural network.

Graph embedding refers to calculation and mapping of high-dimensional feature vectors into low-dimensional vectors, which has the advantage of extracting deeper relationships between vectors and reducing the amount of calculation.

Open domain may be understood as public domain, such as non-infringing video domain. Open domain videos refer to videos in many fields including education, medical treatment, science and the like. The method of the present invention is not limited to a certain field.

In order to make the advantages of the technical schemes of the present invention clearer, the present invention will be described in detail below with reference to the drawings and embodiments.

The present invention provides a multimodal heterogeneous feature fusion-based compact video event description method, including following steps, as shown in Fig. 1:

At S11, feature extracting: a long video in an open domain is selected for clipping to obtain a corresponding video frame by utilizing an I3D convolutional network, and a dynamic visual feature F_m in the video frame is extracted to acquire a dynamic visual feature vector; and an audio prosody feature vector in a video segment is extracted by utilizing a VGGish model;

At S12, scene graph feature embedding: scene object information in the video frame is semantically represented to generate a scene graph, an entity code, an attribute code and a relationship code are obtained, and graph embedding is performed on corresponding feature vectors in the entity code, the attribute code and the relationship code through graph convolution to obtain a scene feature embedding vector.

SPECIFICATION

At S13, triple multi-modal loop fusing: the scene graph feature embedding vector, the dynamic visual feature vector and the audio feature vector are taken as inputs and triple multi-modal loop fusion is performed on these vectors, an output vector is normalized along a channel by means of a normalized function, the normalized output vector is regularized by means of the Dropout and the regularized output vector is input into a decoder, and a preparation is made for reconstructing decoder.

At S14, describing reconstruction decoder, including a pre-trained language dictionary decoder and a description reconstruction algorithm: a video dense event is detected by means of the description reconstruction algorithm, the video dense event here being a collection of events of a video frame; a description of the generated video scene event is decoded by the pre-trained language dictionary decoder, where the pre-trained language dictionary decoder refers to a module that utilizes a pre-trained language dictionary to decode the description of the video scene event.

Based on the above method, the present invention further provides a video dense event description device based on multi-modal heterogeneous feature fusion, and Fig. 2 is a configuration program diagram of the device, as shown in Fig. 2, including:

a feature extraction module 101 configured to select a long video in an open domain for clipping to obtain a corresponding video frame, and extract a dynamic visual feature F_m in the video frame to acquire a dynamic visual feature vector; and extract an audio prosody feature vector in a video segment by utilizing a VGGish model.

Furthermore, the feature extraction module 101 is specifically configured to select a long video in an open domain for clipping to obtain a corresponding video frame by utilizing an I3D convolutional network, and extract a dynamic visual feature F_m in the video frame to obtain a dynamic visual feature vector; for the dynamic visual feature F_m , $F_m = [f_1^m, f_2^m, \dots, f_N^m]$, where N is the number of extracted video clips, and the whole corresponding dimension is $T_v \times d_v$; the dimension corresponding to the audio prosody feature vector is $T_a \times d_a$, and the audio feature vector is converted into a 128-dimensional feature vector with semantics. The network of the VGGish model is essentially a VGG network in the convolutional neural network in structure. The model takes audio data as an input and extracts feature vectors with semantic information from the audio data.

A scene graph feature embedding module 102 configured to semantically represent scene object information in the video frame, generate a scene graph, obtain an entity code, an attribute code and a relationship code, and perform graph embedding on corresponding feature vectors in

SPECIFICATION

the entity code, the attribute code and the relationship code through graph convolution to obtain a scene feature embedding vector.

Furthermore, in the scene graph feature embedding module 102, the input is a video frame, firstly, object position information about object bounding box coordinates, etc. in a video frame picture is detected by means of the Mask R-CNN, a position of an object is determined to obtain an object region; then an original picture of the video frame and the object region are respectively sent to a visual scene graph generator to generate a scene graph, and visual information such as information about the video frame picture and information about the detected object position, etc. are semantically represented to obtain an entity code, an attribute code and a relationship code, these codes are then sent to a joint embedding layer for joint representation to obtain semantic information, the semantic information is embedded into corresponding feature vectors in the entity code, the attribute code and the relationship code; later, graph embedding is performed on the feature vectors by means of graph convolution to obtain a scene graph feature embedding vector; and the association between entities is reinforced inferred in this process.

A triple multi-modal loop fusion module 103 configured to take the scene graph feature embedding vector, the dynamic visual feature vector and the audio feature vector as inputs and perform triple multi-modal loop fusion on these vectors, normalize the output vector along the channel by means of a normalized function, the normalized output vector is regularized by means of the Dropout and the regularized output vector is input into a decoder, and preparation is made for reconstructing decoder.

Furthermore, the triple multi-modal loop fusion module includes a pre-established fusion method considering both parameter quantity and interaction intensity, which enables multiple features to be deeply fused on the premise of not generating high-dimensional features. The normalized function refers to a Softmax function. In which, the dynamic visual feature vector and the audio feature vector with dimensions $T_v \times d_v$ and $T_a \times d_a$ respectively are first sent to a self-attention layer for feature representation.

It should be noted here that the present invention further provides a fusion method considering both parameter quantity and interaction intensity, where the fusion method is configured to realize the triple multi-modal loop fusion in the above method, and enable multiple features to be deeply fused on the premise of not generating high-dimensional features.

Definitely, other multiple fusion are also involved herein, for example, three feature vectors or two feature vectors and one feature graph may be fused at the same time, including loop fusion

SPECIFICATION

performed on any number of modal features on the premise of not considering computer performance. However, as the number of feature fusion increases, it will inevitably lead to the decline of computing performance, and may also lead to the over-fitting problem of the model. Therefore, the fusion effect of three features is good and acceptable.

A description reconstruction decoder module 104 configured to detect a video dense event by means of the description reconstruction algorithm; and decode a description of the generated video scene event by means of the pre-trained language dictionary decoder, where the pre-trained language dictionary decoder refers to a module that utilizes a pre-trained language dictionary to decode the description of the video scene event.

In addition to the functional module in Fig. 2 of the above embodiment, the present invention further provides another video dense event description device based on multi-modal heterogeneous feature fusion, including an adaptive multi-modal data balancing module. Fig. 4 is a configuration program diagram of another video dense event description device based on multi-modal heterogeneous feature fusion provided by the present invention, including a feature extraction module 201, a scene graph feature embedding module 202, a triple multi-modal loop fusion module 203, an adaptive multi-modal data balancing module 204 and a description reconstruction decoder module 205. The adaptive multi-modal data balancing module 204 is configured to extract features of the dynamic visual feature vector and the audio prosody feature vector in a time sequence dimension by utilizing a fully convolutional network, and predict time sequence boundaries and confidence scores of events by using specific values; then, send the predicted events to a common pool, and rank the events in a descending order and select events with high confidence scores; and finally cluster the events with high confidence scores, and select all corresponding clustering centers to form a final event set.

Fig. 3 is a schematic step diagram of another multimodal heterogeneous feature fusion-based compact video event description method provided by the present invention based on the program function module device in Fig. 4, in which the steps of another multimodal heterogeneous feature fusion-based compact video event description method of the present invention are described. The specific steps are as follows:

At S21, feature extraction: a long video in an open domain is clipped to obtain a corresponding video frame and a dynamic visual feature F_m in the video frame is extracted to acquire a dynamic visual feature vector; and an audio prosody feature vector in a video segment is extracted by utilizing a VGGish model.

At S22, scene graph feature embedding: scene object information in the video frame is

SPECIFICATION

semantically represented to generate a scene graph, an entity code, an attribute code and a relationship code are obtained, and graph embedding is performed on corresponding feature vectors in the entity code, the attribute code and the relationship code by using graph convolution to obtain a scene feature embedding vector.

At S23, triple multi-modal loop fusion: the scene graph feature embedding vector, the dynamic visual feature vector and the audio feature vector are taken as inputs and triple multi-modal loop fusion is performed on these vectors, an output vector is normalized along a channel by means of a normalized function, the normalized output vector is regularized by means of the Dropout and the regularized output vector is input into a decoder, and a preparation is made for reconstructing decoder.

At S24, adaptive multi-modal data balancing: features of the dynamic visual feature vector and the audio prosody feature vector are extracted in a time sequence dimension by utilizing a fully convolutional network, and time sequence boundaries and confidence scores of events are predicted with specific values; next, the predicted events are sent to a common pool, and the events are ranked in a descending order and events with high confidence scores are selected; and finally, the events with high confidence scores are clustered, and various corresponding clustering centers are selected to form a final event set.

At S25, description reconstruction decoder, including a pre-trained language dictionary decoder and a description reconstruction algorithm: a video dense event is detected by means of the description reconstruction algorithm; and a generated description of a video scene event is decoded by the pre-trained language dictionary decoder, where the pre-trained language dictionary decoder refers to a module that utilizes a pre-trained language dictionary to decode the description of the video scene event.

In order to understand the specific implementation process of the present invention more clearly, on the basis of the steps of another multimodal heterogeneous feature fusion-based compact video event description method provided in Fig. 3, a specific implementation flowchart of another multimodal heterogeneous feature fusion-based compact video event description method including S24 adaptive multi-modal data balancing is provided herein. As shown in Fig. 5, the present invention may realize the description of dense events in a video stream through the following steps, effectively solving the defect that meaningless descriptions such as co-occurrence descriptions, etc, in a video description method are frequently generated, and effectively utilizing multi-modal information to mine the relationship between scene events, Wherein after a long video in an open domain in an input video stream is clipped in S21, a dynamic visual feature and an

SPECIFICATION

audio prosody feature are extracted respectively; scene graph feature embedding is performed on the extracted dynamic visual feature in S22 to obtain a scene graph feature; triple multi-modal loop feature fusion is performed on three features in S23, and after adaptive multi-modal data balancing in S24, a description of the generated video scene event is obtained through the description reconstruction decoder in S25.

The specific step process is as follows:

At S21, feature extraction, including: a dynamic visual feature is extracted, and a certain number of video are selected for clipping by utilizing an I3D convolutional network, which means that selected long videos in an open domain or an actual surveillance video are clipped one by one; a dynamic visual feature F_m is extracted, smooth flow information is acquired while acquiring the dynamic visual feature, where the smooth flow information here refers to optical flow information. I3D is a combination of a C3D model and a dual-stream network, which not only extends 2D convolution to 3D convolution by adding an extra time dimension to all convolution kernels and pooling kernels in a 2D network, but also adopts an Inception (INC) structure to improve the computational efficiency.

At S21, feature extraction, further including audio prosody information extraction: an audio prosody feature vector is extracted from a video source by utilizing a pre-trained VGGish model. Specifically, the audio prosody feature vector is transformed into a 128-dimensional feature vector with semantics by utilizing a VGGish model, and the 128-dimensional feature vector and the dynamic visual feature vector are fused as the input of a downstream model. Firstly, an input audio is resampled to a 16,000 Hz mono audio; secondly, the audio is subjected to short-time Fourier transform to obtain a spectrogram by utilizing a Hanning time window of 20ms and a frame shift of 15ms; then, the spectrogram is mapped to a 64-order Mel filter bank to calculate a Mel sound spectrum; and finally, these features are framed with a duration of 0.96s, each frame containing 64 Mel frequency bands with a duration of 10ms (that is, 96 frames in total). The model finally outputs a 128-dimensional feature vector, that is, $d_a=128$.

In this embodiment, the I3D convolution network model is as follows: the input video and an optical flow graph of the video are processed in parallel by two branches in the I3D model, and the optical flow graph is obtained by extraction through the opencv, and the results of the two branches are linearly added to obtain the final visual feature as an output. The I3D model used herein pre-trains on a Kinetics data set, and inputs 64 frames of RGB video and optical stream video, each frame size being 224*224, and the video frame rate being 25. In the process of extraction, the input video is cut into a size of 224*224 at the center position, and segmented

SPECIFICATION

according to 64 frames. The I3D model outputs 1024-dimensional RGB feature vector and optical stream feature vector every 64 frames, and finally adds the two vectors to obtain a 1024-dimensional output vector for a 64-frame video segment.

At S22, scene graph feature embedding: scene object information is semantically represented to generate a scene graph, an entity code, an attribute code and a relationship code are obtained, and graph embedding is performed on feature vectors through graph convolution. The scene objects here refer to the objects in the video frame, such as “man/person”, “horse”, “hospital”, “watch” and other entity/physical objects. For details, see the scene graph drawn in Fig. 7. Specifically, object position information about object bounding box coordinates, etc. in a video frame picture is detected by means of the Mask R-CNN, a position of an object is determined to obtain an object region; an original picture of the video frame and the object region are respectively sent to a visual scene graph generator to generate a scene graph, visual information such as information about the video frame picture and information about the detected object position are semantically represented to obtain an entity code, an attribute code and a relationship code, these codes are sent to a joint embedding layer for joint representation to obtain semantic information, the semantic information is embedded into corresponding feature vectors in the entity code, the attribute code and the relationship code; graph embedding is performed on the feature vectors by means of graph convolution to obtain a scene graph feature embedding vector; and the association between entities is reinforced inferred in this process. For the corresponding feature vectors in the entity code, attribute code and relationship code mentioned here, as shown in Fig. 7, the entities are man/person, horse and country road, the relationships are riding and on, and the attributes are red; here is an example for understanding. In practice, these vectors are all encoded vectors.

At S23, triple multi-modal loop fusion: the scene graph feature embedding vector, the dynamic visual feature vector and the audio feature vector are taken as inputs and triple multi-modal loop fusion is performed on these vectors, an output vector is normalized along a channel by means of a normalized function, the normalized output vector is regularized by means of the Dropout and the regularized output vector is input into a decoder, and preparation is made for reconstructing decoder.

Furthermore, firstly, the scene graph feature embedding vector, the dynamic visual feature vector and the audio feature vector are projected into the same low-dimensional space, and loop matrix operation is performed on the feature vectors of the same dimension so as to convert the feature vectors into loop matrices. Starting from the second row of vectors, each row vector is

SPECIFICATION

defined as that the previous row of vectors will circularly move one element to the right on a whole. Secondly, an element product of each vector and all loop matrices except this feature information is made, so as to explore the complex relationships among all the feature vector information. Finally, the feature vectors after interaction of three different modes are normalized and regularized, and then sent to the decoder, and a preparation is made for reconstruction decoder.

At S24, self-adaptive multi-modal data balancing: the dynamic vision feature and the audio prosody feature are fitted with each other to ensure the rationality of event extraction. In one embodiment, to be specific, features of the dynamic visual feature vector and the audio feature vector are extracted in a time sequence dimension by utilizing a fully convolutional network firstly, and time sequence boundaries and confidence scores of events are predicted with three specific values. Secondly, the obtained events are sent to a common pool, and the events are ranked in a descending order and events with high confidence scores are selected. Finally, these events are clustered, and various cluster centers are selected to form a final event set. For example, if a segment of audio feature is full of noise, it is considered here that the audio prosody feature vector in this time dimension seriously affects the dynamic visual feature vector in this dimension, leading to the ineffectiveness of multi-modal fusion and affecting the description of later events. Multi-modal balancing is mainly embodied here, so the dynamic video feature and the audio feature are fitted with each other to ensure the rationality of event extraction.

Furthermore, the three feature values used here to predict time sequence boundaries and confidence scores of the events are as follows:

a center position of the event time period $center = p + \sigma(c)$, where $\sigma(\cdot)$ is a *sigmoid* function configured to scale a value of c within a range of $[0, 1]$ to determine a right neighborhood of a center point p ; an event duration $length = anchor \cdot e$, where e is a scaling factor; and a confidence score $confidence = \sigma(o)$. Here, the confidence score is adopted to screen reasonable events and eliminate meaningless events, for example, during the blank period of the same event, the description of such event cannot be generated.

In this embodiment, the multi-modal data balancing is essentially a fully convolutional network consisting of three convolutional layers. The convolutional network adopts one-dimensional convolution to extract features in the time sequence dimension so as to determine which time periods contain important events. The size of a convolution kernel of the first layer is k , the sizes of both of convolution kernels of the second layer and the third layer are 1, the activation function of each of the first two layers is Relu, and the activation function of the last

SPECIFICATION

layer is *sigmoid*. The module inputs the dynamic visual feature vector and the audio prosody feature vector with dimensions of $T_v \times d_v$ and $T_a \times d_a$ respectively which are output by a coding module. The input of this module is output by the encoding module, with the dynamic visual feature vector $T_v \times d_v$ and the audio prosody feature vector $T_a \times d_a$, respectively. Here, the mechanism is adopted to extract the events in the time sequence dimension for the dynamic visual and audio lines respectively, then the extracted events are input into the common pool for ranking, and an appropriate event set is selected finally.

Furthermore, the convolution operation is specifically as follows: full convolution operations are required to be performed on the dynamic visual feature vector and the audio prosody feature vector respectively for K_v and K_a ($K_v = K_a$) times, so that events with different durations may be captured dynamically, and the k value of the convolution kernel of the first layer is different in each operation so as to ensure that the convolution kernel has rich receptive fields.

Furthermore, the specific number of the predicted events is as follows: here, φ is defined to determine the number of events to be predicted during the m^{th} full convolution operation. Because the dimension T_v of the dynamic visual feature vector is different from the dimension T_a of the audio prosody feature vector, φ_v and φ_a ($\varphi_v \neq \varphi_a$) are respectively defined herein to ensure that $T_v \cdot |\varphi_v|$ and $T_a \cdot |\varphi_a|$ are as close as possible, so that the number of events generated for the two lines are matched as much as possible.

Furthermore, the final event set is specifically as follows: the events extracted in the time sequence dimension are sent to a common pool, all events are ranked in a descending order in the common pool and events with the top 100 confidence scores are selected, finally these events are clustered by means of a K-Means algorithm based on Euclidean distance, and various clustering center are selected to form a final event set.

At S25, description reconstruction decoder, including a pre-trained language dictionary decoder and a description reconstruction algorithm, wherein a pre-training method for the language dictionary includes following steps: converting annotated description text into a scene graph, performing feature embedding by means of a graph convolutional network, calculating the obtained vector and the pre-trained language dictionary D (D is a matrix with a dimension of $d \times L$), re-coding a feature, sending a re-coded vector to a decoder for reconstruction and then for text description, and supervising model learning by comparing the difference between the reconstructed description and the annotated description text. The calculation process of the

SPECIFICATION

re-coding $\varphi(\cdot)$ is as follows:

$$\hat{x} = \varphi(x, D) = \sum_{l=1}^L softmax(D^T x_l) d_l$$

wherein d_l represents the vector in the column 1 of D , and D herein has the function of encoding and embedding human language habits.

Furthermore, in step S25, the description reconstruction algorithm can integrate the generated multiple event descriptions into a complete description so as to ensure the logical timing correctness of the description.

The description reconstruction algorithm specifically includes:

Description Reconstruction Algorithm

Input: start time P and end time Q of a video; event set E , each event containing attributes: central time point m of the event, event duration l , confidence score s , and description z corresponding to the event

Output: candidate description set Z

1. According to m and l , calculating start time s and end time e of each event:

$$s = m - \frac{l}{2}, e = m + \frac{l}{2};$$

2. Calculating tIOU between events, and de-duplicating overlapping events;

3. Ranking the events in an ascending order of start time and selecting an event with the earliest start time as a reference event k ;

4. Selecting events with start time within $[k_s, k_s + \mathcal{E}]$ from a remaining event set $E - k$ to form a start event set E_1 together;

5. Selecting events with start time within $[E_{1e} - \mathcal{E}, E_{1e} + \mathcal{E}]$ from a remaining event set $E - E_1$ to form an event sequence set $\hat{E}$ together with corresponding events in E_1 ;

6. Repeating step 5 and updating $\hat{E}$ until the end time of the selected events is within the range of $[Q - \mathcal{E}, Q]$, updating $\hat{E}$, and ending selection;

7. Merging text descriptions corresponding to every event sequence in $\hat{E}$ and adding the merged text description to a candidate description set Z .

In order to better understand the specific implementation process of scene graph feature embedding in step S22 designed by the present invention, Fig. 6 is listed here, which is a schematic

SPECIFICATION

diagram showing scene graph embedding and generation of another multimodal heterogeneous feature fusion-based compact video event description method provided by the present invention.

As shown in Fig. 6, positions of objects in the figure are detected by means of the Mask R-CNN, and a scene graph G is generated according to the local information about the objects and global information about the objects, where $G = \langle e_o, e_a, e_r \rangle$, and e_o, e_a, e_r respectively represent corresponding vector representations of entities, attributes and relationships in the scene graph. In Fig. 6, a_{ij} represents the j^{th} attribute of the i^{th} entity, o_i represents the i^{th} entity, r_{ij} represents a relationship between o_i and o_j , an entity at the head of an arrow represents a relationship subject, and an entity at the tail represents a relationship object. After the scene graph is generated, graph embedding is performed on entity, relationship and attribute vectors by utilizing four spatial graph convolution layers g_r, g_a, g_s, g_o . The four graph convolution layers all have the same structure: a fully connected layer activated by a Relu function, which may be formally expressed as $g(\cdot) = \text{ReLU}(fc(\text{concat}(e_1, e_2, \dots, e_n)))$, which e_i represents all the input vectors. The relationship feature embedding $x_{r_{ij}}$ is defined as follows: for each pair of relationships $\langle o_i, r_{ij}, o_j \rangle$ in the scene graph,

$$x_{r_{ij}} = g_r(e_{o_i}, e_{r_{ij}}, e_{o_j}).$$

Context relationships of o_i and o_j may be combined by the above formula. The attribute feature embedding x_{a_i} is defined as follows: for a certain entity o_i and an attribute a_{ij} of the entity in the scene graph,

$$x_{a_i} = \frac{\sum_{k=1}^j g_a(e_{o_i}, e_{a_{ik}})}{j}.$$

All attributes of o_i and o_i can be combined by the above formula. The entity feature embedding x_{o_i} is defined as follows: for each entity o_i in the scene graph, because the entity may be both a relationship subject and a relationship object in a certain relationship, Thus, different situations need to be included.

$$x_{o_i} = \frac{1}{N_{r_i}} \left[\sum_{o_j \in \text{sbj}(o_i)} g_s(e_{o_i}, e_{r_{ij}}, e_{o_j}) + \sum_{o_k \in \text{obj}(o_i)} g_s(e_{o_i}, e_{r_{ik}}, e_{o_k}) \right]$$

where sbj refers to a relationship object set taking o_i as relationship subjects, obj refers to a relationship subject set taking o_i as relationship objects, and N_{r_i} refers to the total number of

SPECIFICATION

relationships related to o_i . The entity, attribute and relationship vectors generated in the scene graph are embedded respectively, and the association among adjacent nodes in the graph is inferred by utilizing the graph convolutional network during the embedding process to strengthen the semantic relationships and semantic constraints among nodes, namely associated inference processes.

In order to better understand the specific implementation process of the multi-modal loop fusion in step S3 designed by the present invention, Fig. 7 is listed here, which is a schematic diagram of a triple multi-modal loop fusion module of another multimodal heterogeneous feature fusion-based compact video event description method provided by the present invention.

As shown in Fig. 7, c_T^q and c_T^a are a dynamic visual feature vector and an audio prosody feature vector respectively, and v_T is a feature graph, so projection is performed based on v_T , as shown in the following formulas:

$$V_q = c_T^q W_q^T$$

$$V_a = c_T^a W_a^T$$

$$V_v = W_v v_T$$

where superscript T of the matrix W_i , $i \in \{1,2,3\}$ represents a transpose operation, $W_q \in \mathbb{R}^{d \times a}$, $W_a \in \mathbb{R}^{d \times b}$, $W_v \in \mathbb{R}^c$, and mapped feature vectors $V_q \in \mathbb{R}^d$, $V_a \in \mathbb{R}^d$, $V_v \in \mathbb{R}^d$. At the moment, the feature vectors V_q , V_a and V_v are all in the same dimension, and then loop matrix transformation is performed on these feature vectors respectively. The specific method is as follows: starting from the second row of vectors, each row vector is defined as that the previous row of vectors will circularly move one element to the right on a whole. The transformed loop matrices are shown in the following formulas:

$$C_q = \text{Ciro}(V_q)$$

$$C_a = \text{Ciro}(V_a)$$

$$C_v = \text{Ciro}(V_v)$$

where $C_q \in \mathbb{R}^{d \times d}$, $C_a \in \mathbb{R}^{d \times d}$ and $C_v \in \mathbb{R}^{d \times d}$ are transformed matrices respectively, and $\text{Ciro}(V)$ is a loop matrix transformation function. Finally, Using every kind of feature vectors finishes element product separately with the loop matrix transformed from the other two different vectors, so as to explore the full interaction among the three modal information, as shown in the following formulas:

SPECIFICATION

$$F_q = \frac{1}{d} \sum_{i=1}^d f_i^q \odot C_a \odot C_v$$

$$F_a = \frac{1}{d} \sum_{i=1}^d f_i^a \odot C_q \odot C_v$$

$$F_v = \frac{1}{d} \sum_{i=1}^d f_i^v \odot C_q \odot C_a$$

where f_i is the i^{th} row of feature vectors in the loop matrix corresponding to the modal features, $\odot$ is an element product operation, and $F_q \in \mathbb{R}^d$, $F_a \in \mathbb{R}^d$ and $F_v \in \mathbb{R}^d$ are fused feature vectors.

In this embodiment, the present invention further provides a schematic diagram showing language dictionary training of another multimodal heterogeneous feature fusion-based compact video event description method.

As shown in Fig. 8, in order to prevent the generated description from staying in the superficial co-occurrence description, for example, there are two men in Fig. 8, but the relationship and activities of these two men cannot be described; it is necessary to refine the prior knowledge of language into a knowledge dictionary, and the dictionary can be fused with visual input information for calculation and can guide the model to give a meaningful description with deeper understanding. In order to acquire the dictionary, a self-supervised agent task is designed in this example: an annotated text description S is converted into a scene graph G for the purpose of extracting entities, attributes and relationships from text. Then, the obtained vector and the pre-trained language dictionary D (D is a matrix with a dimension of $d \times L$) are calculated, a feature is re-coded ψ , then a re-coded vector is sent to the decoder for reconstruction and then for text description, and the model learning is supervised by comparing the difference between the reconstructed description and the annotated description. The calculation process of the re-coding $\varphi(\cdot)$ is as follows:

$$\hat{x} = \varphi(x, D) = \sum_{l=1}^L \text{softmax}(D^T x_l) d_l$$

where d_l represents the vector in the column l of D , and D here has the function of encoding and embedding human language habits. Training the language knowledge extraction

SPECIFICATION

model is training the parameters in D actually.

A loss function used to compare the difference between the reconstructed description and the annotated description to supervise the model learning is as follows:

$$loss = \mathbb{E}_{S^s \sim P(s)}[diff(S^s; S)]$$

where S^s represents a description result of the model output, which is a sentence, S is an artificially annotated data set label, and $diff$ represents an algorithm to measure the gap between the model output and the label, which is a CIDEr-D measurement algorithm used in this example.

The present invention further provides an electronic device. The device includes a processor and a memory, wherein the memory stores a configuration program system capable of running the device of the present invention on the processor, wherein the configuration program, when executed by the processor, can implement the video dense event description method of the present invention.

The present invention further provides a computer readable storage medium, wherein the computer readable storage medium stores a configuration program of the device of the present invention, and the configuration program is executable by one or more processors to implement the video dense event description method of the present invention.

The above are only the preferred implementations of the present invention, and it should be noted that for those having ordinary skill in the technical field, several improvements and embellishments can be made without departing from the principle of the present invention, all of which should also be considered to fall within the protection scope of the present invention.

CLAIMS

What is claimed is:

1. A multimodal heterogeneous feature fusion-based compact video event description method, wherein comprises following steps of:

feature extracting: selecting a long video in an open domain for clipping to obtain a corresponding video frame by utilizing an I3D convolutional network, and extracting a dynamic visual feature F_m in the video frame to acquire a dynamic visual feature vector; and extracting an audio prosody feature vector in a video segment by utilizing a VGGish model;

scene graph feature embedding: semantically representing scene object information in the video frame, generating a scene graph, obtaining an entity code, an attribute code and a relationship code, and performing graph embedding on corresponding feature vectors in the entity code, the attribute code and the relationship code through graph convolution to obtain a scene graph feature embedding vector;

triple multi-modal loop fusing: taking the scene graph feature embedding vector, the dynamic visual feature vector and the audio feature vector as inputs and performing triple multi-modal loop fusion on these vectors, then normalizing the output vector along the channel using the normalized function, and using Dropout to regularize and the regularized output vector is input into a decoder to prepare for reconstructing decoder.

describing reconstruction decoder: including a pre-trained language dictionary decoder and a description reconstruction algorithm, detecting a video dense event by means of the description reconstruction algorithm; and decoding a description of the generated video scene event by the pre-trained language dictionary decoder, wherein the pre-trained language dictionary decoder refers to a module that utilizes a pre-trained language dictionary to decode the description of the video scene event.

2. The video dense event description method of claim 1, wherein

for the dynamic visual feature F_m , $F_m = [f_1^m, f_2^m, \dots, f_N^m]$, where N is the number of extracted video clips, and the overall corresponding dimension is $T_v \times d_v$;

the dimension corresponding to the audio prosody feature vector is $T_a \times d_a$, and the audio feature vector is converted into a 128-dimensional feature vector with semantics.

CLAIMS

3. The video dense event description method of claim 1 or 2, wherein

in the process of scene graph feature embedding, the video frame is an input, firstly, object position information such as object bounding box coordinates in a video frame picture is detected by means of the Mask R-CNN first, a position where an object is located is determined to obtain an object region; then an original picture of the video frame and the object region are respectively sent to a visual scene graph generator to generate a scene graph, visual information such as information about the video frame picture and information about the detected object position are semantically represented to obtain an entity code, an attribute code and a relationship code, these codes are then sent to a joint embedding layer for joint representation to obtain semantic information, the semantic information is embedded into corresponding feature vectors in the entity code, the attribute code and the relationship code; later graph embedding is performed on the feature vectors by means of graph convolution to obtain a scene graph feature embedding vector; and the association between entities is reinforced inferred in this process.

4. The video dense event description method of claim 1, wherein

after the triple multi-modal loop fusing, the method further comprises:

adaptive multi-modal data balancing: extracting features of the dynamic visual feature vector and the audio prosody feature vector in a time sequence dimension by utilizing a fully convolutional network, and using specific values to predict time sequence boundaries and confidence scores of events; sending the events obtained through prediction to a common pool, and ranking the events in a descending order and selecting events with high confidence scores; and clustering the events with high confidence scores, and selecting various corresponding clustering centers to form a final event set.

5. The video dense event description method of claim 4, wherein

the specific values refer to a center position $center = p + \sigma(c)$, where $\sigma(\cdot)$ is a *sigmoid* function configured to scale a value of c to a range of $[0, 1]$ to determine a right neighborhood of a center point p ; an event duration $length = anchor \cdot e$, where e is a scaling coefficient; and a confidence score $confidence = \sigma(o)$.

6. The dense event description method of claim 1 or 4 or 5, wherein

CLAIMS

the fully convolutional network adopts one-dimensional convolution to extract features in the time sequence dimension, the size of a convolution kernel of the first layer is k , the sizes of both of convolution kernels of the second layer and the third layer are 1, the activation function of the first two layers is Relu, and the activation function of the last layer is *sigmoid*.

7. The video dense event description method of claim 1 or 4, wherein

in the process of triple multi-modal loop fusing, the normalized function refers to a Softmax function.

8. The video dense event description method of claim 1 or 4, wherein in the process of triple multi-modal loop fusion, the dynamic visual feature vector and the audio feature vector with dimensions of $T_v \times d_v$ and $T_a \times d_a$ respectively are first sent to a self-attention layer for feature representation.

9. The video dense event description method of claim 4, wherein events with the top 100 high confidence scores are selected, and the K-Means algorithm based on Euclidean distance is adopted as a clustering algorithm.

10. The video dense event description method of claim 1, wherein a pre-training method for the language dictionary comprises the following steps of: converting annotated description text into a scene graph, then performing scene graph feature embedding by means of a graph convolutional network, calculating the obtained vector being computed with the pre-trained language dictionary to re-code the scene graph feature, then sending the re-coded vector to a decoder for reconstruction and then performing text description, and supervising model learning by comparing the difference between the reconstructed text description and the annotated description text, the annotated description text being data for pre-training stored in a data set;

the pre-trained language dictionary refers to that the pre-trained language dictionary D is a matrix with a dimension of $d \times L$;

the re-coding refers to the calculation of the re-coding $\varphi(\cdot)$, and the process is as follows:

$$\hat{x} = \varphi(x, D) = \sum_{l=1}^L \text{softmax}(D^T x_l) d_l$$

where d_l represents the vector in the column l of D , and D herein has the function

CLAIMS

of encoding and embedding human language habits.

11. The video dense event description method of claim 1 or 10, wherein in the step of description reconstruction decoder, the description reconstruction algorithm specifically comprises:

Description Reconstruction Algorithm

Input: start time P and end time Q of a video; event set E , each event containing attributes: central time point m of the event, event duration l , confidence score s , and description z corresponding to the event

Output: candidate description set Z

1. According to m and l , calculating start time s and end time e of each event:

$$s = m - \frac{l}{2}, e = m + \frac{l}{2};$$

2. Calculating tIOU between events, and de-duplicating overlapping events;

3. Ranking the events in an ascending order of start time and selecting an event with the earliest start time as a reference event k ;

4. Selecting events with start time within $[k_s, k_s + \mathcal{E}]$ from a remaining event set $E - k$ to form a start event set E_1 together;

5. Selecting events with start time within $[E_{1e} - \mathcal{E}, E_{1e} + \mathcal{E}]$ from a remaining event set $E - E_1$ to form an event sequence set $\hat{E}$ together with corresponding events in E_1 ;

6. Repeating step 5 and updating $\hat{E}$ until the end time of the selected events is within the range of $[Q - \mathcal{E}, Q]$, updating $\hat{E}$, and ending selection;

7. Merging text descriptions corresponding to every event sequence in $\hat{E}$ and adding the merged text description to a candidate description set Z .

12. A fusion method considering both parameter quantity and interaction intensity, wherein the fusion method is configured to realize the triple multi-modal loop fusion described in claim 1, and enable multiple features to be deeply fused on the premise of not generating high-dimensional features.

13. A video dense event description device based on multi-modal heterogeneous feature fusion, comprising:

a feature extraction module, configured to select a long video in an open domain for

CLAIMS

clipping to obtain a corresponding video frame by utilizing an I3D convolutional network, and extract a dynamic visual feature F_m in the video frame to acquire a dynamic visual feature vector; and extract an audio prosody feature vector in a video segment by utilizing a VGGish model;

a scene graph feature embedding module, configured to semantically represent scene object information in the video frame, generate a scene graph, obtain an entity code, an attribute code and a relationship code, and perform graph embedding on corresponding feature vectors in the entity code, the attribute code and the relationship code through graph convolution to obtain a scene feature embedding vector;

a triple multi-modal loop fusion module, configured to take the scene graph feature embedding vector, the dynamic visual feature vector and the audio feature vector as inputs and perform triple multi-modal loop fusion on these vectors, then normalize the output vector along the channel using the normalized function, and use Dropout to regularize the normalized output vector and input the regularized output vector into a decoder to prepare for reconstruction decoder, wherein the triple multi-modal loop fusion module includes a pre-established fusion method considering both parameter quantity and interaction intensity, which enables multiple features to be deeply fused on the premise of not generating high-dimensional features; and

a description reconstruction decoder module, configured to detect a video dense event by means of a description reconstruction algorithm, and decode a description of the generated video scene event by a pre-trained language dictionary decoder, wherein the pre-trained language dictionary decoder refers to a module that utilizes a pre-trained language dictionary to decode the description of the video scene event.

14. The video dense event description device of claim 13, further comprising:

an adaptive multi-modal data balancing module, configured to extract features of the dynamic visual feature vector and the audio prosody feature vector in a time sequence dimension by utilizing a fully convolutional network, and predict time sequence boundaries and confidence scores of events by means of specific values; secondly, send the events obtained through prediction to a common pool, and rank the events in a descending order and select events with high confidence scores; and cluster the events with high confidence scores, and select various corresponding clustering centers to form a final event set.

CLAIMS

15. An electronic device, comprising a processor and a memory storing a configuration program system capable of running the device of claims 13-14 on the processor, wherein the configuration program, when executed by the processor, can implement the video dense event description methods of claims 1-11.

16. A computer readable storage medium, wherein the computer readable storage medium stores a configuration program of the device of claims 13-14, and the configuration program is executable by one or more processors so as to implement the video dense event description methods of claims 1-11.