

A VIDEO ACTION RECOGNITION METHOD AND APPARATUS

Technical field

The present application relates to the field of computer vision, and in particular, to a video action recognition method and a video action recognition apparatus based on sequential-parallel network.

Background

Over the past decade, deep learning models have achieved rapid development and can perform various visual tasks, such as image classification, object detection, and segmentation. The deep learning models have powerful modeling capabilities, and therefore are introduced into videos to perform action recognition tasks. Unlike a 2D image, the unique timeline of the video greatly increases calculation costs, is time-consuming and consumes resources.

However, the increase in the amount of calculation does not mean that time relationship modeling can be completed. Therefore, how to effectively extract appropriate informational spatiotemporal features from the video has become the focus of research. In video action recognition, there are two mainstream architectures. One architecture is based on a dual-stream network. One branch with frame input is used to extract static spatial features, and the other branch with optical flow is used to extract motion features. However, the optical flow branch needs to pre-calculate the optical flow, and therefore, a large computational burden is generated.

The other architecture is based on a three-dimensional (3D) convolutional network. The amount of calculation is reduced by using spatial convolution and temporal convolution alone, but a large amount of calculation still exists. Therefore, there is still much room for improvement on how to extract spatiotemporal features more efficiently.

Summary

Embodiments of the present application provide a video action recognition method, a video action recognition apparatus, a video action recognition device and a computer-readable storage medium.

According to a first aspect, an embodiment of the present application provides a video action recognition method, including:

obtaining a video;

inputting the video into a hidden layer of a neural network model for processing to obtain a recognition object and a motion corresponding to the recognition object of the input video, where the hidden layer of the neural network model includes a plurality of processing units; and

outputting an action recognition result of the input video based on the recognition object in the video and the motion corresponding to the recognition object;

wherein

in at least one processing unit, sequentially identifying spatial features and temporal features of the input video as input of the at least one processing unit, concatenating the spatial features and the temporal features of the input video in parallel, performing a pointwise convolution operation, and outputting spatial semantics and temporal semantics of the input video, wherein the spatial semantics are used to represent the recognition object, and the temporal semantics are used to represent the motion corresponding to the recognition object.

According to the technical scheme provided by the present application, the video processing process has a greater ability to encode spatio-temporal relationships, allowing more meaningful features to be extracted with a smaller number of parameters, thus allowing more useful information to be learned from the dataset using a more compact structure. In the implementation of video processing, only a single processor is required to process the amount of video that previously required multi-processor processing, and the hardware configuration is more optimized, resulting in a more streamlined internal structure of the computer system.

In a possible implementation of the first aspect, the sequentially identifying spatial features and temporal features of the input video includes: performing 2D convolution on the video on spatial axis to obtain the spatial features, to learn the spatial semantics of the input video; and performing 1D convolution on the video on time axis based on the spatial semantics to obtain the temporal features, to learn the temporal semantics of the input video.

In a possible implementation of the first aspect, the concatenating the spatial features and the temporal features of the input video in parallel includes: accumulating channels over the total number of channels of the spatial feature and the channels of the temporal feature, to obtain a concatenated feature map, the spatial feature and the temporal feature are independent of each other; and calculating all channels of the concatenated feature map to obtain values of the concatenated feature map, so as to integrate the spatial and temporal features.

In a possible implementation of the first aspect, the neural network model includes an input layer, multiple hidden layers, and an output layer, the hidden layer further includes a first convolutional layer, a fully convolutional layer, and a global average pooling layer, and the fully convolutional layer includes at least one processing unit, where the input video includes multiple consecutive video frames, and the consecutive video frames pass through the first convolutional layer that includes 24 or 32 output channels and on which three-dimensional convolution operation is performed, to obtain 24 or 32 output channels of consecutive feature maps of the consecutive video frames; the consecutive feature maps of the consecutive video frames are processed by the fully convolutional layer to obtain the recognition object of the input video and the motion corresponding to the recognition object; and the recognition object of the input video and the motion corresponding to the recognition object are processed by the global average pooling layer and are pooled as the action recognition result of the input video.

Compared with other neural network models used in video action recognition, the neural network model uses only half of the amount of calculation to obtain better accuracy, that is, the number of parameters is reduced by 75%, and the amount of calculation is reduced by 50%. Compared with the prior art, According to the technical scheme provided by the present application, the video processing process has a greater ability to encode spatio-temporal relationships, allowing more meaningful features to be extracted with a smaller number of parameters, thus allowing more useful information to be learned from the dataset using a more compact structure. In the implementation of video processing, only a single processor is required to process the amount of video that previously required multi-processor processing, and the hardware configuration is more optimized, resulting in a more streamlined internal structure of the computer system.

In a possible implementation of the first aspect, the neural network model is trained based on something-something v1 and v2 datasets that do not require pre-training, so that the SPNet model has low latency and high throughput.

According to a second aspect, an embodiment of the present application provides a video action recognition apparatus, comprising: obtaining module, obtaining a video; processing module, processing the input video based a neural network model to obtain a recognition object and a motion corresponding to the recognition object of the input video, wherein the processing module comprises a plurality of processing units and; outputting module, outputting an action recognition result of the input video based on the recognition object in the video and the motion corresponding to the recognition object; wherein in at least one processing unit, comprising:

sequential processing sub-module, identifying spatial features and temporal features of the input video as input of the at least one processing unit; parallel processing sub-module, concatenating the spatial features and the temporal features of the input video in parallel, and performing a pointwise convolution operation; and outputting sub-module, outputting spatial semantics and temporal semantics of the input video, wherein the spatial semantics are used to represent the recognition object, and the temporal semantics are used to represent the motion corresponding to the recognition object.

In a possible implementation of the second aspect, the sequential processing sub-module comprises: performing 2D convolution on the video on spatial axis to obtain the spatial features, to learn the spatial semantics of the input video; and performing 1D convolution on the video on time axis based on the spatial semantics to obtain the temporal features, to learn the temporal semantics of the input video.

In a possible implementation of the second aspect, the parallel processing sub-module comprises: accumulating channels over the total number of channels of the spatial feature and the channels of the temporal feature, to obtain a concatenated feature map, the spatial feature and the temporal feature are independent of each other; and calculating all channels of the concatenated feature map to obtain values of the concatenated feature map, so as to integrate the spatial and temporal features.

In a possible implementation of the second aspect, the apparatus further includes that the neural network model comprises an input layer, multiple hidden layers, and an output layer,

the hidden layer further comprises a first convolutional layer, a fully convolutional layer, and a global average pooling layer, and

the fully convolutional layer comprises at least one processing unit,

Wherein, the input video comprises multiple consecutive video frames, and the consecutive video frames pass through the first convolutional layer that comprises 24 or 32 output channels and on which three-dimensional convolution operation is performed, to obtain 24 or 32 output channels of consecutive feature maps of the consecutive video frames;

the consecutive feature maps of the consecutive video frames are processed by the fully convolutional layer to obtain the recognition object of the input video and the motion corresponding to the recognition object; and

the recognition object of the input video and the motion corresponding to the recognition object are processed by the global average pooling layer and are pooled as the action recognition result of the input video.

In a possible implementation of the second aspect, the apparatus further includes that the neural network model is trained based on something-something v1 and v2 datasets for which pre-training is not needed.

According to a third aspect, an embodiment of the present application provides a video action recognition device, including:

a memory, configured to store instructions executed by one or more processors of a system, and

a processor, being one of the processors of the system, and configured to execute the instructions to implement the video action recognition method, wherein:

obtaining a video;

inputting the video into a hidden layer of a neural network model for processing, so as to obtain a recognition object and a motion corresponding to the recognition object of the input video, wherein the hidden layer of the neural network model comprises a plurality of processing units, and;

outputting an action recognition result of the input video based on the recognition object in the video and the motion corresponding to the recognition object;

wherein

in at least one processing unit, sequentially identifying spatial features and temporal

features of the input video as input of the at least one processing unit, concatenating the spatial features and the temporal features of the input video in parallel, and performing a pointwise convolution operation, and outputting spatial semantics and temporal semantics of the input video, wherein the spatial semantics are used to represent the recognition object, and the temporal semantics are used to represent the motion corresponding to the recognition object.

In a possible implementation of the third aspect, the device further includes that the sequentially identifying spatial features and temporal features of the input video, comprises:

performing 2D convolution on the video on spatial axis to obtain the spatial features, to learn the spatial semantics of the input video; and

based on the spatial semantics, performing 1D convolution on the video on time axis to obtain the temporal features, to learn the temporal semantics of the input video.

In a possible implementation of the third aspect, the device further includes that the concatenating the spatial features and the temporal features of the input video in parallel, comprises:

accumulating channels over the total number of channels of the spatial feature and the channels of the temporal feature, to obtain a concatenated feature map, the spatial feature and the temporal feature are independent of each other; and

calculating all channels of the concatenated feature map to obtain values of the concatenated feature map, so as to integrate the spatial and temporal features.

In a possible implementation of the third aspect, the device further includes that the neural network model comprises an input layer, multiple hidden layers, and an output layer,

the hidden layer further comprises a first convolutional layer, a fully convolutional layer, and a global average pooling layer, and

the fully convolutional layer comprises at least one processing unit,

wherein, the input video comprises multiple consecutive video frames, and the consecutive video frames pass through the first convolutional layer that comprises 24 or 32 output channels and on which three-dimensional convolution operation is performed, to obtain 24 or 32 output channels of consecutive feature maps of the consecutive video frames;

the consecutive feature maps of the consecutive video frames are processed by the fully convolutional layer to obtain the recognition object of the input video and the motion corresponding to the recognition object; and

the recognition object of the input video and the motion corresponding to the recognition object are processed by the global average pooling layer and are pooled as the action recognition result of the input video.

In a possible implementation of the third aspect, the device further includes that the neural network model is trained based on something-something v1 and v2 datasets for which pre-training is not needed.

According to a forth aspect, an embodiment of the present application provides a computer-readable storage medium encoded with a computer program, where instructions are stored on the computer-readable medium, and when the instructions are executed on a computer, the computer performs the video action recognition method according to any one of possible implementations of the first aspect.

Brief description of the drawings

FIG. 1 is a schematic structural diagram of a neural network model according to some embodiments of the present application;

FIG. 2 is an architectural schematic structural diagram of a neural network model according to some embodiments of the present application;

FIG. 3 is a structural block diagram of an electronic device according to some embodiments of the present application;

FIG. 4 is a flowchart of a video action recognition method according to some embodiments of the present application;

FIG. 5(a) is a schematic structural diagram of an R(2+1)D structure sequentially connecting spatial convolution and temporal convolution according to some embodiments of the present application;

FIG. 5(b) is a schematic structural diagram of a GST structure connecting spatial convolution and temporal convolution in parallel according to some embodiments of the present application;

FIG. 5(c) is a schematic structural diagram of a processing unit according to some embodiments of the present application;

FIG. 6(a) is a network structural diagram of a processing unit according to some embodiments of the present application;

FIG. 6(b) is a network structural diagram of a C3D structure according to some embodiments of the present application;

FIG. 6(c) is a network structural diagram of a P3D-C structure according to some embodiments of the present application;

FIG. 6(d) is a network structural diagram of only a parallel structure according to some embodiments of the present application;

FIG. 6(e) is a network structural diagram of only a sequential structure according to some embodiments of the present application;

FIG. 7 is a schematic structural diagram of a video action recognition apparatus according to some embodiments of the present application;

FIG. 8 is a diagram of comparison between performance of a neural network model and performance of other video action recognition models according to some embodiments of the present application;

FIG. 9 is a schematic structural diagram of an electronic device for video action recognition according to some embodiments of the present application; and

FIG. 10 is a schematic structural diagram of an electronic device according to some embodiments of the present application.

Detailed description of the embodiments

The illustrative embodiments of the present application include, but are not limited to, a video action recognition method, a video action recognition apparatus, a video action recognition device and a computer-readable storage medium.

It can be understood that, as used herein, the terms "module" and "unit" may refer to or include an application-specific integrated circuit (ASIC), an electronic circuit, a processor (shared, dedicated, or group) that executes one or more software or firmware programs and/or a memory, a combinational logic circuit, and/or other suitable hardware components that provide the described functions, or may be used as a part of these hardware components.

It can be understood that in the embodiments of the present application, a processor may be a microprocessor, a digital signal processor, a microcontroller, etc., and/or any combination thereof. According to another aspect, the processor may be a single-core processor, a multi-core processor, etc., and/or any combination thereof.

It can be understood that the video action recognition method provided in the present application can be implemented on various electronic devices, including but not limited to electronic devices such as a server, a distributed server cluster consisting of multiple servers, a mobile phone, a tablet computer, a laptop computer, a desktop computer, a wearable device, a head-mounted display, a mobile email device, a portable game console, a portable music player, a reader device, a personal digital assistant, a virtual reality or augmented reality device, and a television with one or more processors embedded or coupled.

In particular, the video action recognition method provided by the present application is suitable for an edge device. Edge computing adjacently provides edge intelligence services by using a distributed open platform (architecture) that integrates core capabilities of network, computing, storage, and application on the edge of the network near the source of objects or data, so that key requirements of real-time service, data optimization, application intelligence, security and privacy protection can be met. For example, the edge device may be a device capable of performing edge calculation on video data at an end close to a video data source (a network intelligent camera) in a video surveillance system.

As shown in Figure 1, according to some embodiments of the present application, a neural

network model is provided, and is called an SPNet model, which belongs to the fully convolutional network.

The neural network model shown in Figure 1 can be used for video action recognition. As shown in Figure 1, the neural network model consists of input layer, multiple implicit layers, and output layer for a total of n network layers. Of these, the first layer is called the input layer, the last layer is called the output layer, and the other layers are called the hidden layer. The input layer receives objects to be processed in different dimensions in the form of arrays, such as pictures, video frames, etc. The hidden layer consists of three types of underlying constructs: convolutional, pooling, and fully connected layers. In the basic construction, the convolutional and pooled layers are specific to convolutional neural networks. More complex algorithms include complex constructions such as Inception modules, residual blocks, etc. The convolutional cores in the convolutional layer contain weighting factors, while the pooling layer does not.

As shown in Figure 1, the structure of the underlying network model consists of 1 input layer, 3 convolutional layers, 2 pooling layers, 1 fully connected layer and 1 output layer. The order of the three types of foundation construction in the implicit layer is usually: input layer - convolutional layer - pooling layer - fully connected layer - output layer.

A specific process of processing the neural network model shown in Figure 1 by using a sample image when video action recognition is performed is described in detail below.

Before the neural network model is used to perform video action recognition, a large amount of sample data and expected result data need to be inputted into the model for training. A video is divided into several consecutive video frames, each video frame is used as a sample image, and image data of each sample image is inputted to n nodes of the input layer of the neural network model shown in Figure 1. Calculation of the hidden layer is performed, and video action recognition result data is finally generated after calculation of the output layer. After training of each image is completed, the processing result data finally outputted by the neural network model is compared with the expected result data, an error is calculated, and weights of nodes in network layers other than the input layer are adjusted based on the error. In this way, the neural network model is trained by inputting data of the consecutive video frames of the video, and the weights are continuously adjusted. When the error between the facial video action recognition result data finally outputted by the neural network model and the expected result data is less than an error threshold, it is determined that the neural network model converges.

Figure 2 is a schematic architectural structural diagram of a neural network model according to some embodiments of the present application. As shown in Figure 2, ResNet-50 consists of the first convolutional layer, the full convolutional layer, and the pooling layer, in which the full connection layer consists of four stages, corresponding to four groups of blocks, each of which consists of 3, 4, 6, and 3 blocks in turn, called residual blocks. Each block has three convolutional layers, thus forming the 50-layer ResNet network model.

The architecture of the SPNet network model is similar to that of ResNet-50, but there are also significant differences between the two. The SPNet model uses a 50-layer network structure similar to that of ResNet-50, consisting of a first convolutional layer, a full convolutional layer and a pooling layer, where the full convolutional layer consists of multiple stages, each consisting of several processing units. The difference is that the full convolutional layer of the ResNet-50 consists of residual blocks, whereas the full convolutional layer of the SPNet model consists of processing units according to the present application. In said processing unit, sequentially identifying spatial features and temporal features of the input video as input of the processing unit, concatenating the spatial features and the temporal features of the input video in parallel, and performing a pointwise convolution operation, and outputting spatial semantics and temporal semantics of the input video as input data, wherein the spatial semantics are used to represent the recognition object, and the temporal semantics are used to represent the motion corresponding to the recognition object.

Figure 2 illustrates an architectural schematic of a neural network model in accordance with some embodiments of the present application. As shown in Figure 2, T RGB frames were sampled from the video and used as input. The first convolution layer is a simple three-dimensional convolution that is pooled through a global average pooling layer after the last processing unit, and then connected to a full connection layer to classify the input video for prediction. Each stage contains several processing units, and after each stage, the number of channels output doubles.

In the present application, the number of output channels of the first convolutional layer is expressed as the hyperparameter B. When $B = x$, this neural network model is named as SPNet-Bx. e.g., when $B = 24$, this neural network model is named as SPNet-B24.

The hyperparameter B represents the number of output channels of the first convolutional layer and is defined as a precision control parameter for network-level complexity. Specifically to this neural network model, as shown in Figure 2, the hyperparameter B represents the number

of filters in the first processing unit, and as the processing phase progresses, the number of filters based on the hyperparameter B increases at a 2-fold rate, namely 2B, 4B, 8B, 16B.

In the present application, it was found, that when a value of the hyperparameter B is set to 24 and 32, the neural network model becomes compact and efficient. Therefore, the SPNet model is instantiated, and a specific network structure of the SPNet model is shown in Table 1. Values of d_2 , d_3 , d_4 , and d_5 respectively represent the number of output channels of a previous processing unit. It is set that a step of 2 is applied to the spatial convolutional layer, a time step is not used, and spatial down-sampling occurs in the first processing unit of each stage. After each convolution operation, BatchNorm and ReLU operations are used.

Table 1

Stage	Output size	SPNet-B24	SPNet-B32
data layer	$T \times 224^2$	-	-
conv1	$T \times 112^2$	$3, 3 \times 7 \times 7, 24$	$3, 3 \times 7 \times 7, 32$
stage2	$T \times 56^2$	$\begin{bmatrix} d_2, 1 \times 1^2, 24 \\ 24, 1 \times 3^2, 24 \\ 24, 3 \times 1^2, 24 \\ 48, 1 \times 1^2, 48 \end{bmatrix} \times 3$	$\begin{bmatrix} d_2, 1 \times 1^2, 32 \\ 32, 1 \times 3^2, 32 \\ 32, 3 \times 1^2, 32 \\ 64, 1 \times 1^2, 64 \end{bmatrix} \times 3$
stage3	$T \times 28^2$	$\begin{bmatrix} d_3, 1 \times 1^2, 48 \\ 48, 1 \times 3^2, 48 \\ 48, 3 \times 1^2, 48 \\ 96, 1 \times 1^2, 96 \end{bmatrix} \times 4$	$\begin{bmatrix} d_3, 1 \times 1^2, 64 \\ 64, 1 \times 3^2, 64 \\ 64, 3 \times 1^2, 64 \\ 128, 1 \times 1^2, 128 \end{bmatrix} \times 4$
stage4	$T \times 14^2$	$\begin{bmatrix} d_4, 1 \times 1^2, 96 \\ 96, 1 \times 3^2, 96 \\ 96, 3 \times 1^2, 96 \\ 192, 1 \times 1^2, 192 \end{bmatrix} \times 6$	$\begin{bmatrix} d_4, 1 \times 1^2, 128 \\ 128, 1 \times 3^2, 128 \\ 128, 3 \times 1^2, 128 \\ 256, 1 \times 1^2, 256 \end{bmatrix} \times 6$
stage5	$T \times 7^2$	$\begin{bmatrix} d_5, 1 \times 1^2, 192 \\ 192, 1 \times 3^2, 192 \\ 192, 3 \times 1^2, 192 \\ 384, 1 \times 1^2, 384 \end{bmatrix} \times 3$	$\begin{bmatrix} d_5, 1 \times 1^2, 256 \\ 256, 1 \times 3^2, 256 \\ 256, 3 \times 1^2, 256 \\ 512, 1 \times 1^2, 512 \end{bmatrix} \times 3$
pool	$1 \times 1 \times 1$	global average pool	
	# classes	fully connected layer	

Compared with other neural network models used in video action recognition, the neural network model uses only half of the amount of calculation to obtain better accuracy, that is, the number of parameters is reduced by 75% while the amount of computation is reduced by 50%, in the implementation of video processing, only a single processor is needed to process the amount of video that previously required multi-processor processing, so that the hardware configuration is more optimized, and the internal structure of the computer system is further streamlined.

It can be understood that the above description of the technical solution for training the neural network model shown in Figure 1 is merely exemplary and not limiting.

Figure 3 is a structural block diagram of an electronic device 100 according to some embodiments of the present application. Specifically, as shown in Figure 3, the electronic device 100 includes one or more processors 104, system control logic 108 connected to at least one of the processors 104, a system memory 112 connected to the system control logic 108, and a nonvolatile memory (NVM) 116 connected to the system control logic 108, and a network interface 120 connected to the system control logic 108.

In some embodiments, the processor 104 may include one or more single-core or multi-core processors. In some embodiments, the processor 104 may include any combination of a general-purpose processor and a dedicated processor (for example, a graphics processing unit, an application processor, and a baseband processor). In an embodiment in which the electronic device 100 uses an evolved NodeB (Evolved Node B, eNB) or radio access network (radio access network, RAN) controller, the processor 104 may be configured to perform various corresponding embodiments.

In some embodiments, the processor 104 may be configured to perform video action recognition based on a neural network model. Specifically, for example, in a process in which an image is processed by each layer of the neural network model, the processor 104 continuously adjusts the neural network model according to acquired image information until a highly accurate image feature map is obtained. The processor 104 may further periodically update the neural network model to better adapt to changes in various actual demands on the neural network model.

In some embodiments, the system control logic 108 may include any suitable interface controller to provide any suitable interface for at least one of the processors 104 and/or any suitable device or component in communication with the system control logic 108.

In some embodiments, the system control logic 108 may include one or more memory controllers to provide an interface connected to the system memory 112. The system memory 112 may be used to load and store data and/or instructions. In some embodiments, the memory 112 of the electronic device 100 may include any suitable volatile memory, such as a suitable dynamic random access memory (DRAM). In some embodiments, the system memory 112 may be used to load or store instructions for implementing the above video action recognition, or the

system memory 112 may be used to load or store instructions for implementing an application program to perform video action recognition by using the neural network model.

The NVM/memory 116 may include one or more tangible, non-transitory computer-readable media for storing data and/or instructions. In some embodiments, the NVM/memory 116 may include any suitable nonvolatile memory such as a flash memory and/or any suitable nonvolatile storage device, such as at least one of a hard disk drive (Hard Disk Drive, HDD), a compact disc (Compact Disc, CD) drive, and a digital versatile disc (Digital Versatile Disc, DVD) drive. The NVM/memory 116 can be further configured to store the neural network model used in the above video action recognition.

The NVM/memory 116 may include a part of storage resources on an apparatus on which the electronic device 100 is installed, or may be accessed by the device, but is not necessarily a part of the device. For example, the NVM/storage 116 can be accessed by using the network via the network interface 120.

In particular, the system memory 112 and the NVM/memory 116 may respectively include: a temporary copy and a permanent copy of an instruction 124. The instruction 124 may include an instruction that the electronic device 100 is caused to implement the video action recognition method shown in this application when the instruction is executed by at least one of the processors 104. In some embodiments, the instruction 124, and hardware, firmware, and/or software components thereof may additionally/alternatively be placed in the system control logic 108, the network interface 120, and/or the processor 104.

The network interface 120 may include a transceiver for providing a radio interface for the electronic device 100, to further communicate with any other suitable devices (such as a front-end module or an antenna) by using one or more networks. In some embodiments, the network interface 120 may be integrated with other components of the electronic device 100. For example, the network interface 120 may be integrated into at least one of the processor 104, the system memory 112, the NVM/memory 116, and a firmware device (not shown) with an instruction, and when at least one of the processors 104 executes the instruction, the electronic device 100 implements a video action recognition method shown in this application.

The network interface 120 may further include any suitable hardware and/or firmware to provide a multiple-input multiple-output radio interface. For example, the network interface 120 may be a network adapter, a wireless network adapter, a telephone modem, and/or a wireless

modem.

In some embodiments, at least one of the processors 104 may be packaged with logic for one or more controllers of the system control logic 108 to form a system in package (SiP). In some embodiments, at least one of the processors 104 may be integrated on the same die with the logic for one or more controllers of the system control logic 108 to form a system on chip (SoC).

The electronic device 100 may further include: an input/output (I/O) device 132. The I/O device 132 may include a user interface, so that a user can interact with the electronic device 100. A peripheral component can also interact with the electronic device 100 by using the design of a peripheral component interface. In some embodiments, the electronic device 100 further includes a sensor for determining at least one of environmental conditions and location information related to the electronic device 100.

According to some embodiments of the present application, a video action recognition method 400 is provided. Figure 4 is a flowchart of a video action recognition method according to some embodiments of the present application. As shown in Figure 4, the video action recognition method 400 performed by using an SPNet model includes the following steps:

Step 402: Obtain a video

Step 404: The input video is processed into multiple consecutive video frames at the input layer of the SPNet mode.

Step 406: Consecutive video frames pass through the first convolutional layer that includes 24 or 32 output channels and on which three-dimensional convolution operation is performed, to obtain 24 or 32 output channels of consecutive feature maps of the consecutive video frames.

Step 408: The consecutive feature maps of the consecutive video frames are processed by a fully convolutional layer to obtain a recognition object of the input video and a motion corresponding to the recognition object.

Step 410: Being processed by the global average pooling layer, the recognition object of the input video and the motion corresponding to the recognition object are pooled as an action recognition result of the input video.

Step 412: The action recognition result of the input video is outputted by the process of the output layer.

In some embodiments, a hidden layer of the SPNet model includes a plurality of processing units (that is also called sequential-parallel blocks in the present application), and at least one or more sequential-parallel blocks among the plurality of sequential-parallel blocks are interconnected and connected to the output of the input layer.

In some embodiments, in at least one processing unit, spatial features and temporal features of the input video are sequentially identified as input of the at least one processing unit, the spatial features and the temporal features of the input video are concatenated in parallel, a pointwise convolution operation is performed, and spatial semantics and temporal semantics of the input video are outputted. The spatial semantics are used to represent the recognition object, and the temporal semantics are used to represent the motion corresponding to the recognition object.

In other embodiments, a fully connected layer may be added at the end of the process, to predict the classification of each category.

Compared with the existing video action recognition method, there is a stronger capability to encode a spatiotemporal relationship, and more meaningful features can be extracted by using a smaller number of parameters in the present video action recognition method. In this way, a more compact structure can be used, a model capacity can be reduced, and more useful information can be learned from a dataset.

Next, with reference to FIGS. 5(a) through 5(c), a detailed description of the improvements made by the processing unit of the present application relative to the prior art.

In the structure shown in Figure 5(a) to 5(c), the size of the convolutional nucleus is represented by $T \times H \times W$, where T represents the time size, and H , W represent the spatial height and width of the convolutional nucleus, respectively.

Figure 5(a) is a schematic diagram of an R(2+1)D structure sequentially connecting spatial convolution and temporal convolution according to some embodiments of the present application. As shown in Figure 5(a), R(2+1)D is the most common way to sequentially connect the two convolutions, as shown in Figure 5(a) $1 \times 3 \times 3$ for spatial convolution and $3 \times 1 \times 1$ for temporal convolution.

Figure 5(b) is a schematic diagram of a GST structure connecting spatial convolution and temporal convolution in parallel according to some embodiments of the present application. As

shown in Figure 5(b), for the GST structure, the input features are (optionally) divided into two parts before spatial convolution and spatial-temporal convolution are performed. However, because spatial features are repeatedly extracted in two branches in a GST module, feature extraction is redundant.

Figure 5(c) is a schematic structural diagram of a processing unit according to some embodiments of the present application. As shown in Figure 5(b), the GST structure divides the input features into two parts before performing spatial and temporal convolution, and then performs $1 \times 3 \times 3$ spatial convolution and $3 \times 3 \times 3$ temporal convolution operations for these two parts, respectively. The above operation will cause spatial features to be extracted repeatedly in both branches of the GST module, resulting in redundant feature extraction.

FIG. 5(c) illustrates a structural schematic of a processing unit according to some embodiments of the present application, and as shown in FIG. 5(c), the processing performed in the processing unit can be divided into two stages, one of which is a sequential stage and the other is a parallel stage.

For the sequential phase of the processing unit, it consists of two decomposed 2D convolution $1 \times 3 \times 3$ and 1D convolution $3 \times 1 \times 1$, performing 2D convolution operations on the spatial axis to obtain spatial features and 1D convolution operations on the time axis to obtain temporal features, respectively. Spatial convolution learns the spatial semantics of the input video clip, and temporal convolution learns the temporal semantics of the input video clip. The spatial convolution operates before the temporal convolution to extract meaningful spatial features from different video frames, and the temporal convolution looks for motion patterns that match the spatial semantics with an understanding of the spatial semantics.

For the parallel phase of the process in the processing unit, spatial and temporal features flow into the parallel phase in parallel, and a concatenate operation (concatenate) is performed before the point-by-point convolution is performed. In contrast, in the $R(2 + 1)D$ block of the prior art, the spatio-temporal features extracted by the sequentially connected convolution flow directly to the $1 \times 1 \times 1$ point-by-point convolution. Compared to sequential connections alone, the processing unit can be operated by a combination of concatenate operations and point-by-point convolutional operations that enhance the interaction between spatio-temporal features without introducing additional computations. Because spatial features are reused rather than redundantly extracted in the processing unit, it makes the processing unit more efficient.

The instantiation of the sequential-parallel block is presented in Table 2. As shown in Table 2, a block-level complexity-accuracy control parameter γ is introduced, and is used to control the ratio between the number of input channels to the spatial convolution and the final number of output channels of the block. The value of the hyperparameter γ is typically 1/2 in the experiments unless otherwise specified. Combined with the previous hyperparameter B, a series of networks that have different levels of computation and spatial-temporal modeling capability and cater to different needs and application scenarios can be obtained.

Table 2:

Stage	Input channels	Kernel size	Output channels
-	N	$1 \times 1 \times 1$	$\gamma \times N$
Sequential	$\gamma \times N$	$1 \times 3 \times 3$	$\gamma \times N$
	$\gamma \times N$	$3 \times 1 \times 1$	$\gamma \times N$
Parallel	$2 \times \gamma \times N$	$1 \times 1 \times 1$	N

In some embodiments, the sequentially identifying spatial features and temporal features of the video includes, performing 2D convolution on the video on spatial axis to obtain the spatial features, to learn the spatial semantics of the video, and performing 1D convolution on the video on time axis based on the spatial semantics to obtain the temporal features, to learn the temporal semantics of the video. Specifically, recognition of an action of a human hand is used as an example. First, the spatial semantics are obtained by learning the video on the spatial axis, that is, the recognition object is the human hand, and then, on the premise that the recognition object is the human hand, temporal semantics learning is performed on the video on the temporal axis, that is, recognizing the action of the hand, to determine whether the action is the motion such as waving, shaking or washing.

The following specifically explains the concatenate operation of spatial and temporal features in video data. Concatenate processing of spatial and temporal features in video data according to the present application comprises: accumulating channels over the total number of channels of the spatial feature and the channels of the temporal feature, to obtain a concatenated feature map, the spatial feature and the temporal feature are independent of each other; and calculating all channels of the concatenated feature map to obtain values of the concatenated feature map, so as to integrate the spatial and temporal features.

Specifically, the concatenate operation is the accumulation operation for all channels of the

spatial and temporal features, i.e., the number of channels of the spatial feature and the number of channels of the temporal feature are added together. For example, a spatial feature with an $N/2$ (more specific example, 32) output channel feature map is obtained by performing a two-dimensional convolution on the spatial axis, and a time feature with an $N/2$ (more specific example, 32) output channel feature map is obtained by performing a one-dimensional convolution on the time axis. The concatenate operation of the feature map of the spatial feature and the feature map of the time feature results in a concatenate feature map of N (i.e. $32 + 32 = 64$) output channels. That is, the concatenate operation adds up the total number of channels in the spatial and temporal feature maps and makes the temporal and spatial features independent of each other.

After obtaining separate spatial and temporal features, the two types of feature need to be fused together to obtain spatio-temporal features, which need to be rolled and manipulated point by point. In the process of point-by-point convolution, all channels in a particular region of the feature map are computed simultaneously, outputting values for a particular location in the feature map, thereby fusing the previously combined temporal and spatial features.

In order to illustrate the benefits of concatenate operations according to this application, a comparison of this application and the prior art is described below. FIGS. 6(a) to (e) illustrate a schematic representation of the structure of the processing unit of the present application and the processing unit of the prior art, together with the change in the number of channels during its processing.

All the structures in Figure 6(a), (b), (d), and (e) undergo the same pointwise convolution before and after the processing shown in the figure, where N denotes the number of input channels. The P3D-C structures shown in Figure 6(c) are treated differently from them.

FIG. 6(a) shows a schematic network structure diagram of a structure of a processing unit and a schematic representation of the variation of the number of channels during its processing, according to some embodiments of the present application, as shown in FIG. 6(a), the spatial convolution of the input channels of the processing unit is $1 \times 3 \times 3$ of $N/2$ and the temporal convolution of $3 \times 1 \times 1$ of $N/2$ of the number of channels, and after a concatenate operation, the superposition of the channels is performed to obtain the output result of $1 \times 1 \times 1$ of N of the number of channels.

Figure 6(b) shows a network structure diagram of a C3D structure in an existing technique

where the C3D performs only $1 \times 3 \times 3$ spatial convolution operations with an input channel count of $N/2$.

Figure 6(c) shows a network structure diagram of a P3D-C structure in the prior art, as shown in Figure 6(c), the P3D-C structure is similar to the GST structure shown in Figure 5(b), where the subsequent $1 \times 1 \times 1$ convolution of the input channels is $N / 2$. In P3D-C, the output features of the spatial and temporal convolution are directly added together using summation as a feature fusion, e.g. when the spatial and temporal convolution respectively obtain a feature diagram of 32 output channels, after summation the feature diagram of 32 output channels remains. Whereas, in the processing unit of the present application shown in FIG. 6(a), the features are combined to output a feature map of the 64 output channels. In P3D-C design, the temporal convolution is mainly used as a residual of the spatial convolution, which limits the time modeling capability to some extent. In contrast, the processing units in this application are operated in such a way that spatial and temporal features are preserved intact using association, and then fused using point-by-point convolution, resulting in greater spatio-temporal modeling capabilities for the SPNet model consisting of processing units.

Figure 6(d) shows a network structure diagram for a parallel-only structure in the existing technique, as shown in Figure 6(d), with a spatial convolution of $1 \times 3 \times 3$ for the number of input channels of $N/2$ and a temporal convolution of $3 \times 1 \times 1$ for the number of channels of $N/2$, with no superposition of channels, just the result of parallel output of 2 channels of $N/2$.

Figure 6(e) illustrates a network structure diagram for a sequential-only structure in the prior art, where the number of output channels in the temporal convolution of the sequential structure is N , resulting in an increase in computational cost compared to the processing unit structure shown in Figure 6(a).

In some embodiments, the neural network model is trained based on something-something v1 and v2 datasets for which pre-training is not needed. Specifically, something-something v1 and v2 datasets are two large action classification datasets that require large-scale spatiotemporal modeling, including 108K and 220K videos, respectively. To understand the videos and correctly classify them, they require a high capability of spatio-temporal reasoning because extensive spatial interactions between recognition objects and complex temporal relations exist. Unlike previous models, SPNet models are trained from scratch (random initialization) without performing pre-training such as ImageNet or dynamics. As a result, the SPNet model has low

latency and high throughput.

According to some embodiments of the present application, a video action recognition apparatus 600 is provided. Figure 7 illustrates a structural schematic of a video motion recognition apparatus in accordance with some embodiments of the present application. As shown in FIG. 7, the apparatus 600 for video motion recognition using the SPNet model is as follows:

obtaining module 602, obtaining a video;

processing module 604 , processing the input video based a neural network model to obtain a recognition object and a motion corresponding to the recognition object of the input video, wherein the processing module comprises a plurality of processing units such as processing unit1, processing unit2, processing unit3,etc. and;

outputting module 606, outputting an action recognition result of the input video based on the recognition object in the video and the motion corresponding to the recognition object; wherein in at least one processing unit, (for example, processing unit1)comprising:

sequential processing sub-module 1, identifying spatial features and temporal features of the input video as input of the at least one processing unit;

parallel processing sub-module 1, concatenating the spatial features and the temporal features of the input video in parallel, and performing a pointwise convolution operation; and

outputting sub-module 1, outputting spatial semantics and temporal semantics of the input video, wherein the spatial semantics are used to represent the recognition object, and the temporal semantics are used to represent the motion corresponding to the recognition object.

In some embodiments, the sequential processing sub-module comprises: performing 2D convolution on the video on spatial axis to obtain the spatial features, to learn the spatial semantics of the input video; and performing 1D convolution on the video on time axis based on the spatial semantics to obtain the temporal features, to learn the temporal semantics of the input video.

In some embodiments, the parallel processing sub-module comprises: accumulating channels over the total number of channels of the spatial feature and the channels of the temporal feature, to obtain a concatenated feature map, the spatial feature and the temporal feature are

independent of each other; and calculating all channels of the concatenated feature map to obtain values of the concatenated feature map, so as to integrate the spatial and temporal features.

In some embodiments, the neural network model comprises an input layer, multiple hidden layers, and an output layer,

the hidden layer further comprises a first convolutional layer, a fully convolutional layer, and a global average pooling layer, and

the fully convolutional layer comprises at least one processing unit,

Wherein, the input video comprises multiple consecutive video frames, and the consecutive video frames pass through the first convolutional layer that comprises 24 or 32 output channels and on which three-dimensional convolution operation is performed, to obtain 24 or 32 output channels of consecutive feature maps of the consecutive video frames;

the consecutive feature maps of the consecutive video frames are processed by the fully convolutional layer to obtain the recognition object of the input video and the motion corresponding to the recognition object; and

the recognition object of the input video and the motion corresponding to the recognition object are processed by the global average pooling layer and are pooled as the action recognition result of the input video.

In some embodiments, the neural network model is trained based on something-something v1 and v2 datasets for which pre-training is not needed.

Next, advantages of the SPNet model are described with reference to the experimental data. The input video is divided into T segments, and a center frame of each segment is selected to form a T frame input. For time sampling, the same number of frames (16 or 24) is used as input. For spatial sampling, first, the length of the short side is adjusted to 224, and then the central square with a size 224 is cut out as the input model. Just like temporal sampling, only one spatial crop is used for inference in the technical solution of the present application (unless otherwise specified).

Figure 8 is a diagram of comparison between performance of a neural network model and performance of other video action recognition models according to some embodiments of the present application. The horizontal coordinate represents the number of GFLOPs (Giga

Floating-point Operations) used to measure the computing capability of the computer, and the vertical coordinate represents accuracy. As shown in Figure 7, the SPNet model uses less than 20 GFLOPs to achieve competitive performance, and uses less than 30 GFLOPs to achieve optimal performance. Therefore, it is concluded that SPNet has certain advantages in both accuracy and calculation costs. It should be noted that, unlike other models that use pre-trained ResNet models, SPNet is trained from scratch on the dataset.

Figure 9 is a schematic structural diagram of an electronic device for video action recognition according to some embodiments of the present application. As shown in Figure 9, the electronic device 800 includes:

- a video obtaining module 802, configured to obtain a video;
- an input layer 804, processing the video into multiple consecutive video frames;
- a first convolutional layer 806, processing the multiple consecutive video frames into consecutive feature maps;
- a fully convolutional layer 808, outputting the consecutive feature maps as a recognition object of the video and a motion corresponding to the recognition object;
- a global average pooling layer 810, pooling the recognition object of the video and the motion corresponding to the recognition object as an action recognition result of the video; and
- an output layer 812, outputting the action recognition result of the video.

It can be understood that the electronic device 800 for video action recognition shown in Figure 9 corresponds to the video action recognition method provided by the present application. The technical details in the above specific description of the video action recognition method provided by the present application are still applicable to the electronic device 800 for video action recognition shown in Figure 9. For a detailed description, please refer to the above contents. Details are not described herein.

It can be understood that exemplary applications of video action recognition based on a neural network provided by the embodiments of the present application include but are not limited to video action recognition in the field of artificial intelligence.

Figure 10 is a schematic structural diagram of an electronic device 900 according to an

embodiment of the present application. The electronic device 900 can also execute the video action recognition method disclosed in the above embodiments of the present application. In Figure 10, similar components have the same reference signs. As shown in Figure 10, the electronic device 900 may include a processor 910, a power module 940, a memory 980, a mobile communication module 930, a wireless communication module 920, a sensor module 990, an audio module 950, a camera 970, an interface module 960, a key 901, a display screen 902, and so on.

It can be understood that the structure illustrated in this embodiment of the present application does not constitute a specific limitation on the electronic device 900. In other embodiments of the present application, the electronic device 900 may include more or less components than shown, or some components are combined, or split some components are split, or different components are arranged. The illustrated components can be implemented in hardware, software, or a combination of software and hardware.

The processor 910 may include one or more processing units, for example, may include a processing unit or a processing circuit such as a central processing unit (Central Processing Unit, CPU), a video action recognizer (Graphics Processing Unit, GPU), a digital signal processor (Digital Signal Processor, DSP), a micro-programmed control unit (Micro-Programmed Control Unit, MCU), an artificial intelligence (Artificial Intelligence, AI) processor, or a field programmable gate array (Field Programmable Gate Array, FPGA). Different processing units may be independent devices, or may be integrated in one or more processors. A storage unit is disposed in the processor 910, and is configured to store instructions and data. In some embodiments, the storage unit in processor 910 is a cache memory 980. The memory 980 mainly includes a storage program area 9801 and a data storage area 9802. The storage program area 9801 may store an operating system and application programs required by at least one function (functions such as voice playback or image recognition). The video action recognition method based on the neural network model provided in the embodiment of the present application can be regarded as an application program in the stored program area 9801 to implement functions such as video action recognition. The neural network model is stored in the above data storage area 9802.

The power module 940 may include a power supply, a power management component, and the like. The power supply may be a battery. The power management component is used to manage charging of the power supply and supplying power to other modules by the power

supply. In some embodiments, the power management component includes a charging management module and a power management module. The charging management module is configured to receive charging input from a charger. The power management module is configured to connect the power supply, the charging management module, and the processor 910. The power management module receives input from the power supply and/or the charging management module, and supplies power to the processor 910, the display screen 902, the camera 970, and the wireless communication module 920.

The mobile communication module 930 may include, but is not limited to, an antenna, a power amplifier, a filter, a low noise amplifier (Low Noise Amplifier, LNA), and the like. The mobile communication module 930 may provide a solution of wireless communication including 2G/3G/4G/5G and the like applied to the electronic device 900. The mobile communication module 930 can receive electromagnetic waves from the antenna, filter and amplify the received electromagnetic waves, and transmit the electromagnetic waves to a modem processor for demodulation. The mobile communication module 930 can further amplify the signal modulated by the modem processor, and convert the signal into electromagnetic waves for radiation through the antenna. In some embodiments, at least some of the functional modules of the mobile communication module 930 may be disposed in the processor 910. In some embodiments, at least some of the functional modules of the mobile communication module 930 may be arranged in the same device as at least some of the modules of the processor 910. The wireless communication technology may include global system for mobile communications (global system for mobile communications, GSM), general packet radio service (general packet radio service, GPRS), Code Division Multiple Access (code division multiple access, CDMA), and Wideband Code Division Multiple Access (wideband code division multiple access, WCDMA), Time Division-Synchronous Code Division Multiple Access (Time Division-Synchronous Code Division Multiple Access, TD-SCDMA), Long Term Evolution (Long Term Evolution, LTE), Bluetooth (Bluetooth, BT), global navigation satellite system (global navigation satellite system, GNSS), wireless local area network (wireless local area network, WLAN), near field communication (near field communication, NFC), frequency modulation (frequency modulation, FM) and/or an infrared (infrared, IR) technology, etc. The GNSS may include a global positioning system (global positioning system, GPS), a Global Navigation Satellite System (Global Navigation Satellite System, GLONASS), a BeiDou Navigation Satellite System (BeiDou Navigation Satellite System, BDS), a Quasi-Zenith Satellite System (Quasi-Zenith Satellite System, QZSS), and/or a Satellite Based Augmentation System (Satellite Based

Augmentation System, SBAS).

The wireless communication module 920 may include an antenna, and transmitting and receiving of electromagnetic waves are implemented via the antenna. The wireless communication module 920 can provide a solution of wireless communication including wireless local area network (wireless local area network, WLAN) (such as wireless fidelity (wireless fidelity, Wi-Fi) network), Bluetooth (Bluetooth, BT), and global navigation satellite system (global navigation satellite system, GNSS), frequency modulation (FM), a near field communication (near field communication, NFC) technology, or an infrared (infrared, IR) technology that is applied to the electronic device 900. The electronic device 900 can communicate with the network and other devices by using the wireless communication technology.

In some embodiments, the mobile communication module 930 and the wireless communication module 920 of the electronic device 900 may be located in the same module.

The display screen 902 is used to display human-computer interaction interfaces, images, videos, and the like. The display screen 902 includes a display panel. The display panel may use a liquid crystal display (liquid crystal display, LCD), an organic light-emitting diode (organic light-emitting diode, OLED), or an active matrix organic light-emitting diode (active-matrix organic light emitting diode, AMOLED), a flexible light-emitting diode (flexible light-emitting diode, FLED), Miniled, MicroLed, Micro-oled, a quantum dot light emitting diode (quantum dot light emitting diode, QLED), etc.

The sensor module 990 may include a proximity light sensor, a pressure sensor, a gyro sensor, an air pressure sensor, a magnetic sensor, an acceleration sensor, a distance sensor, a fingerprint sensor, a temperature sensor, a touch sensor, an ambient light sensor, a bone conduction sensor, and the like.

The audio module 950 is configured to convert digital audio information into an analog audio signal for outputting, or convert analog audio input into a digital audio signal. The audio module 950 may further be configured to encode and decode audio signals. In some embodiments, the audio module 950 may be disposed in the processor 910, or some of the functional modules of the audio module 950 may be disposed in the processor 910. In some embodiments, the audio module 950 may include a loudspeaker, an earpiece, a microphone, and a headset jack.

The camera 970 is used to capture static images or videos. The object generates an optical image by using a lens and projects the optical image onto a photosensitive element. The photosensitive element converts an optical signal into an electrical signal, and then transmits the electrical signal to an image signal processor (Image Signal Processor, ISP) to be converted into a digital image signal. The electronic device 900 can implement a shooting function by using an ISP, the camera 970, a video codec, a graphics processing unit (Graphic Processing Unit, GPU), the display screen 902, an application processor, and the like.

The interface module 960 includes an external memory interface, a universal serial bus (universal serial bus, USB) interface, and a subscriber identification module (subscriber identification module, SIM) card interface. The external memory interface may be used to connect an external memory card, such as a Micro SD card, to expand a storage capacity of the electronic device 900. The external memory card communicates with the processor 910 by using the external memory interface to implement a data storage function. The universal serial bus interface is used for the electronic device 900 to communicate with other electronic devices. The user identification module card interface is used to communicate with the SIM card installed in the electronic device 900, for example, to read a phone number stored in a SIM card or write the phone number into the SIM card.

In some embodiments, the electronic device 900 further includes a key 901, a motor, an indicator, and the like. The key 901 may include a volume key, an on/off key, and so on. The motor is used to cause the electronic device 900 to generate a vibration effect. For example, when the electronic device 900 of a user is called, vibration is generated to remind the user of answering an incoming call of the electronic device 900. The indicator may include a laser indicator, a radio frequency indicator, an LED indicator, and the like.

The embodiments of the mechanism disclosed in the present application may be implemented in hardware, software, firmware, or a combination of these implementation methods. The embodiments of the present application may be implemented as a computer program or program code executed on a programmable system. The programmable system includes at least one processor, a storage system (including a volatile and non-volatile memory and/or a storage element), at least one input device, and at least one output device.

The program code can be applied to input instructions to perform the functions described in the present application and generate output information. The output information can be applied to one or more output devices in a known manner. For the objectives of the present application, a

processing system includes any system having a processor such as a digital signal processor (DSP), a microcontroller, an application specific integrated circuit (ASIC), or a microprocessor.

The program code can be implemented in a high-level programming language or an object-oriented programming language to communicate with the processing system. Alternatively, the program code can be implemented in assembly language or machine language if necessary. In fact, the mechanism described in the present application is not limited to the scope of any particular programming language. In either case, the language can be a compilation language or an interpreted language.

In some cases, the disclosed embodiments may be implemented in hardware, firmware, software, or any combination thereof. The disclosed embodiments can further be implemented as instructions carried by or stored on one or more transitory or non-transitory machine-readable (for example, computer-readable) storage media, and the instructions can be read and executed by one or more processors. For example, the instructions may be distributed by using the network or by using other computer-readable media. Therefore, the machine-readable media may include any mechanism for storing or transmitting information in a machine-readable (for example, a computer-readable) form, including but not limited to, a floppy disk, a compact disk, an optical disc, a read-only memory (CD-ROM), a magneto optical disc, a read only memory (ROM), a random access memory (RAM), an erasable programmable read only memory (EPROM), an electrically erasable programmable read only memory (EEPROM), a magnetic card or an optical card, a flash memory, or a tangible machine-readable memory used to transmit information (for example, a carrier, an infrared signal, and a digital signal) in an electrical, optical, or acoustic form, or other forms of propagating signals by using the Internet. Therefore, the machine-readable media include any type of machine-readable media suitable for storing or transmitting electronic instructions or information in a machine-readable (for example, computer-readable) form.

In the drawings, some structural or method features may be shown in a specific arrangement and/or order. However, it should be understood that such a specific arrangement and/or ordering may not be needed. Rather, in some embodiments, these features may be arranged in a manner

and/or order different from the manner and/or order shown in the illustrative drawings. In addition, the inclusion of structural or method features in a particular figure does not imply that such features are needed in all embodiments, and these features may not be included or may be combined with other features in some embodiments.

It should be noted that each unit/module mentioned in each device embodiment of the present application is a logical unit/module. Physically, a logical unit/module may be a physical unit/module or a part of a physical unit/module, and can further be implemented by a combination of multiple physical units/modules. The physical implementation of these logical units/modules is not the most important, but the combination of the functions implemented by these logical units/modules is the key to solving the technical problem raised by the present application. In addition, in order to highlight the innovative part of the present application, the above device embodiments of the present application do not introduce units/modules that are not closely related to solving the technical problem raised by the present application, but this does not mean that the above device embodiments do not include other units/modules.

It should be noted that, in the examples and the specification of this patent, relational terms such as "first" and "second" are used only to distinguish one entity or operation from another entity or operation, and do not necessarily require or imply that there is any such actual relationship or order between these entities or operations. Moreover, the terms "comprise", "include" or any other variant thereof are intended to cover non-exclusive inclusion, so that a process, a method, an article, or a device that includes a series of elements includes not only those elements, but also other elements not explicitly listed, or elements that are inherent to this process, method, article, or device. In the absence of more limitations, the element defined by the sentence "include one" does not exclude that there are other identical elements in the process, method, article or device that includes the element.

Although the present application has been illustrated and described by referring to certain preferred embodiments of the present application, persons of ordinary skill in the art should understand that various changes can be made in form and detail without departing from the spirit

and scope of the present application.

Claims

1. A video action recognition method, comprising:

obtaining a video, the video comprises multiple consecutive video frames;

inputting the video into a hidden layer of a neural network model for processing, so as to obtain a recognition object and a motion corresponding to the recognition object of the input video, wherein the hidden layer of the neural network model comprises a plurality of processing units, and;

outputting an action recognition result of the input video based on the recognition object in the video and the motion corresponding to the recognition object;

wherein

in at least one processing unit, sequentially identifying spatial features and temporal features of the input video as input of the at least one processing unit, concatenating the spatial features and the temporal features of the input video in parallel, and performing a pointwise convolution operation, and outputting spatial semantics and temporal semantics of the input video, wherein the spatial semantics are used to represent the recognition object, and the temporal semantics are used to represent the motion corresponding to the recognition object;

wherein the sequentially identifying spatial features and temporal features of the input video, comprises:

performing 2D convolution on the video on spatial axis to obtain the spatial features, to learn the spatial semantics of the input video; and

based on the spatial semantics, performing 1D convolution on the video on time axis to obtain the temporal features, to learn the temporal semantics of the input video;

wherein the concatenating the spatial features and the temporal features of the input video in parallel, comprises:

accumulating channels over the total number of channels of the spatial feature and the channels of the temporal feature, to obtain a concatenated feature map, the spatial feature and the

temporal feature are independent of each other; and

calculating all channels of the concatenated feature map to obtain values of the concatenated feature map, so as to integrate the spatial and temporal features.

2. The method according to claims 1, wherein the neural network model comprises an input layer, multiple hidden layers, and an output layer,

the hidden layer further comprises a first convolutional layer, a fully convolutional layer, and a global average pooling layer, and

the fully convolutional layer comprises at least one processing unit,

wherein the consecutive video frames pass through the first convolutional layer that comprises 24 or 32 output channels and on which three-dimensional convolution operation is performed, to obtain 24 or 32 output channels of consecutive feature maps of the consecutive video frames;

the consecutive feature maps of the consecutive video frames are processed by the fully convolutional layer to obtain the recognition object of the input video and the motion corresponding to the recognition object; and

the recognition object of the input video and the motion corresponding to the recognition object are processed by the global average pooling layer and are pooled as the action recognition result of the input video.

3. The method according to claim 2, wherein the neural network model is trained based on something-something v1 and v2 datasets for which pre-training is not needed.

4. A video action recognition apparatus, comprising:

obtaining module, obtaining a video, the video comprises multiple consecutive video frames;

processing module, processing the input video based a neural network model to obtain a

recognition object and a motion corresponding to the recognition object of the input video, wherein the processing module comprises a plurality of processing units and;

outputting module, outputting an action recognition result of the input video based on the recognition object in the video and the motion corresponding to the recognition object;

wherein

in at least one processing unit, comprising:

sequential processing sub-module, identifying spatial features and temporal features of the input video as input of the at least one processing unit;

parallel processing sub-module, concatenating the spatial features and the temporal features of the input video in parallel, and performing a pointwise convolution operation; and

outputting sub-module, outputting spatial semantics and temporal semantics of the input video, wherein the spatial semantics are used to represent the recognition object, and the temporal semantics are used to represent the motion corresponding to the recognition object;

wherein the sequential processing sub-module comprises:

performing 2D convolution on the video on spatial axis to obtain the spatial features, to learn the spatial semantics of the input video; and

performing 1D convolution on the video on time axis based on the spatial semantics to obtain the temporal features, to learn the temporal semantics of the input video;

wherein the parallel processing sub-module comprises:

accumulating channels over the total number of channels of the spatial feature and the channels of the temporal feature, to obtain a concatenated feature map, the spatial feature and the temporal feature are independent of each other; and

calculating all channels of the concatenated feature map to obtain values of the concatenated feature map, so as to integrate the spatial and temporal features.

5. The apparatus according to claims 4, wherein the neural network model comprises an input layer, multiple hidden layers, and an output layer,

the hidden layer further comprises a first convolutional layer, a fully convolutional layer, and a global average pooling layer, and

the fully convolutional layer comprises at least one processing unit,

wherein the consecutive video frames pass through the first convolutional layer that comprises 24 or 32 output channels and on which three-dimensional convolution operation is performed, to obtain 24 or 32 output channels of consecutive feature maps of the consecutive video frames;

the consecutive feature maps of the consecutive video frames are processed by the fully convolutional layer to obtain the recognition object of the input video and the motion corresponding to the recognition object; and

the recognition object of the input video and the motion corresponding to the recognition object are processed by the global average pooling layer and are pooled as the action recognition result of the input video.

6. The apparatus according to claim 5, wherein the neural network model is trained based on something-something v1 and v2 datasets for which pre-training is not needed.

7. A video action recognition device, comprising:

a memory, configured to store instructions executed by one or more processors of a system, and

a processor, being one of the processors of the system, and configured to execute the instructions to implement the video action recognition method, wherein:

obtaining a video, the video comprises multiple consecutive video frames;

inputting the video into a hidden layer of a neural network model for processing, so as to obtain a recognition object and a motion corresponding to the recognition object of the input video, wherein the hidden layer of the neural network model comprises a plurality of processing units, and;

outputting an action recognition result of the input video based on the recognition object in the video and the motion corresponding to the recognition object;

wherein

in at least one processing unit, sequentially identifying spatial features and temporal features of the input video as input of the at least one processing unit, concatenating the spatial features and the temporal features of the input video in parallel, and performing a pointwise convolution operation, and outputting spatial semantics and temporal semantics of the input video, wherein the spatial semantics are used to represent the recognition object, and the temporal semantics are used to represent the motion corresponding to the recognition object;

wherein the sequentially identifying spatial features and temporal features of the input video, comprises:

performing 2D convolution on the video on spatial axis to obtain the spatial features, to learn the spatial semantics of the input video; and

based on the spatial semantics, performing 1D convolution on the video on time axis to obtain the temporal features, to learn the temporal semantics of the input video;

wherein the concatenating the spatial features and the temporal features of the input video in parallel, comprises:

accumulating channels over the total number of channels of the spatial feature and the channels of the temporal feature, to obtain a concatenated feature map, the spatial feature and the temporal feature are independent of each other; and

calculating all channels of the concatenated feature map to obtain values of the concatenated feature map, so as to integrate the spatial and temporal features.

8. The device according to claims 7, wherein the neural network model comprises an input layer, multiple hidden layers, and an output layer,

the hidden layer further comprises a first convolutional layer, a fully convolutional layer, and a global average pooling layer, and

the fully convolutional layer comprises at least one processing unit,

wherein the consecutive video frames pass through the first convolutional layer that comprises 24 or 32 output channels and on which three-dimensional convolution operation is performed, to obtain 24 or 32 output channels of consecutive feature maps of the consecutive video frames;

the consecutive feature maps of the consecutive video frames are processed by the fully convolutional layer to obtain the recognition object of the input video and the motion corresponding to the recognition object; and

the recognition object of the input video and the motion corresponding to the recognition object are processed by the global average pooling layer and are pooled as the action recognition result of the input video.

9. The device according to claim 8, wherein the neural network model is trained based on something-something v1 and v2 datasets for which pre-training is not needed.

10. A computer-readable storage medium encoded with a computer program, wherein instructions are stored on the computer-readable medium, and when the instructions are executed on a computer, the computer performs the video action recognition method according to any one of claims 1-3.