

MULTI-MODAL PRE-TRAINING METHOD AND APPARATUS

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] The present application is based on and claims priority from the application of CN application No. 202111078728.2 filed on September 15, 2021, the disclosure of hereby incorporated herein by reference in its entirety.

TECHNICAL FIELD

[0002] The present disclosure relates to the field of information processing, and in particular, to a multi-modal pre-training method and a multi-modal pre-training apparatus.

BACKGROUND

[0003] The multi-modal pre-training technology for visual language is one of recently emerging subjects in the multi-modal field, which aims at enabling a model to pre-train large-scale weakly labeled visual data (such as images and videos) and text data to obtain a better multi-modal feature representation, thereby improving the performance of various multi-modal task models.

[0004] Technologies related to the multi-modal pre-training for the visual language are basically methods of pre-training a model with reference to BERT (Bidirectional Encoder Representations From Transformer) in the field of natural language processing.

SUMMARY

[0005] According to a first aspect of embodiments of the present disclosure, there is provided a multi-modal pre-training method, comprising: sampling a video in a video-text pair to obtain a first video frame sequence; performing word segmentation processing on a text in the video-text pair to obtain a first word segmentation sequence; masking on the first video frame sequence to obtain a second video frame sequence; masking on the first word segmentation sequence to obtain a second word segmentation sequence; encoding the first video frame sequence to obtain a first video feature, and encoding the first word segmentation sequence to obtain a first word segmentation feature; encoding the second video frame sequence to obtain a second video feature, and encoding the second word segmentation sequence to obtain a second word segmentation feature; determining a pre-trained target function by using the first video feature, the first word segmentation feature, the second video feature and the second word segmentation feature; and performing multi-modal pre-training by using the pre-trained target function.

[0006] In some embodiments, the determining a pre-trained target function comprises: determining a first contrastive loss value by using the first word segmentation feature, the second video feature and a preset first negative sample feature; determining a second contrastive loss value by using the first video feature, the second word segmentation feature and a preset second negative sample feature; determining a first target based on the first contrastive loss value and second contrastive loss value; determining a third contrastive loss value by using the first video feature, the second video feature, and the second negative sample feature; determining a fourth contrastive loss

value by using the first word segmentation feature, the second word segmentation feature, and the first negative sample feature; determining a second target according to the third contrastive loss value and the fourth contrastive loss value; and determining the target function according to the first target and the second target.

[0007] In some embodiments, the determining a first contrastive loss value comprises: converting the first word segmentation feature into a global first positive sample feature; converting the second video feature into a global video query feature; and determining a first contrastive loss value by using the video query feature, the first positive sample feature, and the first negative sample feature.

[0008] In some embodiments, the determining a second contrastive loss value comprises: converting the first video feature into a global second positive sample feature; converting the second word segmentation feature into a global text query feature; and determining a second contrastive loss value by using the text query feature, the second positive sample feature, and the second negative sample feature.

[0009] In some embodiments, the determining a third contrastive loss value comprises: determining a third contrastive loss value by using the video query feature, the second positive sample feature, and the second negative sample feature.

[0010] In some embodiments, determining a fourth contrastive loss value comprises: determining a fourth contrastive loss value by using the text query feature, the first positive sample feature, and the first negative sample feature.

[0011] In some embodiments, the first target is a sum of the first contrastive loss value and the second contrastive loss value; and the second target is a sum of the third contrastive loss value and the fourth contrastive loss value.

[0012] In some embodiments, the target function is a sum of the first target and the second target.

[0013] In some embodiments, fusing the second video feature and the second word segmentation feature to obtain a fused feature; and inputting the fused feature into a masked language modelling (MLM) model to obtain a third target, and inputting the fused feature into a masked language generating (MSG) model to obtain a fourth target; and the determining the target function according to the first target and the second target comprising: determining the target function according to the first target, the second target, the third target and the fourth target.

[0014] In some embodiments, the target function is a sum of the first target, the second target, the third target, and the fourth target.

[0015] According to a second aspect of embodiments of the present disclosure, there is provided a multi-modal pre-training apparatus, comprising: a first processing module configured to sample a video in a video-text pair to obtain a first video frame sequence, and perform word segmentation processing on a text in the video-text pair to obtain a first word segmentation sequence; a second processing module configured to mask on the first video frame sequence to obtain a second video frame sequence, and mask on the first word segmentation sequence to obtain a second word segmentation sequence; a third processing module configured to encode the first video frame sequence to obtain a first video feature, and encode the first word segmentation sequence to obtain a first word segmentation feature; a fourth processing module configured to encode the second video frame sequence to obtain a second video feature, and encode the second word segmentation sequence to obtain a second word segmentation feature; a fifth processing module configured to determine a pre-trained target function by using

the first video feature, the first word segmentation feature, the second video feature and the second word segmentation feature; and a sixth processing module configured to perform multi-modal pre-training by using the pre-trained target function.

[0016] According to a third aspect of embodiments of the present disclosure, there is provided a multi-modal pre-training apparatus, comprising: a memory; and a processor coupled to the memory, which is configured to execute the method according to any one of the embodiments described above.

[0017] According to a fourth aspect of embodiments of the present disclosure, there is provided a non-transitory computer-readable storage medium, wherein the computer-readable storage medium stores computer instructions, which, when executed by a processor, implement the method according to any one of the embodiments described above.

[0018] Other features of the present disclosure and advantages thereof will become apparent from the following detailed description of exemplary embodiments thereof, which proceeds with reference to the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[0019] In order to more clearly illustrate the embodiments of the present disclosure or the technical solutions in the prior art, drawings used in the embodiments or the description of the prior art will be briefly described below. It is obvious that the drawings in the following description are only some embodiments of the present disclosure, and for those skilled in the art, other drawings may be obtained according to the drawings without inventive labor.

[0020] FIG. 1 is a schematic flow chart diagram of a multi-modal pre-training method according to one embodiment of the present disclosure;

[0021] FIG. 2 is a schematic flow chart diagram of a multi-modal pre-training method according to another embodiment of the present disclosure;

[0022] FIG. 3 is a schematic structural diagram of a multi-modal pre-training apparatus according to one embodiment of the present disclosure;

[0023] FIG. 4 is a schematic structural diagram of a multi-modal pre-training apparatus according to another embodiment of the present disclosure;

[0024] FIG. 5 is a schematic diagram of a multi-modal pre-training model according to one embodiment of the present disclosure.

DETAILED DESCRIPTION

[0025] The technical solutions in the embodiments of the present disclosure in a clear and complete manner with reference to the figures in the embodiments of the present disclosure, and it is obvious that the embodiments described are only some, rather than all, of the embodiments of the present disclosure. The following description of at least one exemplary embodiment is merely illustrative in nature and is in no way intended to limit the disclosure, its application, or uses. All other embodiments, which can be derived by a person skilled in the art from the embodiments disclosed herein without paying inventive effort, are intended to be within the scope of the present disclosure.

[0026] The relative arrangement of parts and steps, numerical expressions and numerical values set forth in these embodiments do not limit the scope of the present disclosure unless specifically stated otherwise.

[0027] Meanwhile, it should be understood that the sizes of the respective portions shown in the figures are not drawn in an actual proportional relationship for the convenience of description.

[0028] Techniques, methods, and apparatus known to one of ordinary skill in the relevant art may not be discussed in detail but are intended to be part of the specification where appropriate.

[0029] In all examples shown and discussed herein, any particular value should be construed as exemplary only and not as limiting. Thus, other examples of the exemplary embodiments may have different values.

[0030] It should be noted that similar reference numbers and letters refer to similar items in the following figures, and thus, once an item is defined in one figure, it need not be

discussed further in subsequent figures.

[0031] The inventors find that in the related art, in order to mine the connection between two modalities, the video text multi-modal pre-training technology only utilizes a masked input video text to learn the relevance of the global feature representation during the pre-training period, and such a learning manner fails to enable sufficient exploration of the overall video-text relation between the input video frame and the word sequence, thereby causing a degradation in the quality of multi-modal features.

[0032] Accordingly, a multi-modal pre-training scheme is provided by the present disclosure, which can enhance the relevance between cross-modal data and effectively improve the comprehension capability of a multi-modal pre-training model on multi-modal data contents.

[0033] FIG. 1 is a schematic flow chart diagram of a multi-modal pre-training method according to one embodiment of the present disclosure. In some embodiments, the following multi-modal pre-training method is performed by a multi-modal pre-training apparatus.

[0034] In step 101, a video in a video-text pair is sampled to obtain a first video frame sequence, and a word segmentation process is performed on a text in the video-text pair to obtain a first word segmentation sequence.

[0035] In some embodiments, the video is sampled in an equidistant sampling manner to obtain the first video frame sequence.

[0036] In some embodiments, flag [CLS] and flag [SEP] are respectively provided at the beginning and end of the first word segmentation sequence for convenience of subsequent processing.

[0037] In step 102, the first video frame sequence is masked to obtain a second video frame sequence, and the first word

segmentation sequence is masked to obtain a second word segmentation sequence.

[0038] In some embodiments, video frames in the first video frame sequence are replaced with masks at a random probability to obtain a second video frame sequence.

[0039] In some embodiments, word segments in the first word segmentation sequence are replaced with masks at a random probability to obtain a second word segmentation sequence.

[0040] In step 103, the first video frame sequence is encoded to obtain a first video feature, and the first word segmentation sequence is encoded to obtain a first word segmentation feature.

[0041] In some embodiments, the first video frame sequence is encoded by using a Video Key Encoder to obtain a first video feature, and the first word segmentation sequence is encoded by using a Sentence Key Encoder to obtain a first word segmentation feature.

[0042] The first video feature outputted from the video key encoder reflects contextual characteristics of unmasked video frames. The first word segmentation feature outputted from the sentence key encoder reflects contextual characteristics of unmasked word segmentation sequences.

[0043] Since the video key and the sentence key are not inventive focus of the present disclosure, they will not be described in detail here.

[0044] In step 104, the second video frame sequence is encoded to obtain a second video feature, and the second word segmentation sequence is encoded to obtain a second word segmentation feature.

[0045] In some embodiments, the second video frame sequence is encoded by using a Video Query Encoder to obtain a second video feature, and the second word segmentation sequence is encoded by using a Sentence Query Encoder to obtain a second word

segmentation feature.

[0046] The second video feature outputted from the video query encoder reflects a relevance between frames in the video modality, and the second word segmentation feature outputted from the sentence query encoder reflects a relevance between words in the text modality.

[0047] Since the video query encoder and the sentence query encoder are not inventive focus of the present disclosure, they will not be described in detail here.

[0048] In step 105, a pre-trained target function is determined by using the first video feature, the first word segmentation feature, the second video feature and the second word segmentation feature.

[0049] In some embodiments, determining the pre-trained target function is as illustrated in FIG. 2.

[0050] In step 201, a first contrastive loss value is determined by using the first word segmentation feature, the second video feature and a preset first negative sample feature.

[0051] In some embodiments, the first word segmentation feature is converted to a global first positive sample feature H_s^+ by using an MLP (Multi-layer Perceptron) model, and the second video feature is converted to a global video query feature H_v^m by using the MLP model. The first contrastive loss value is determined by using the video query feature H_v^m , the first positive sample feature H_s^+ and a first negative sample feature K_s^- .

[0052] It should be noted that the first negative sample feature K_s^- is represented by the following Formula:

$$K_s^- = \{H_{s,i}^-\}_{i=1}^K \quad (1)$$

[0053] In the Formula (1), K represents a size of a negative sample queue included in the first negative sample feature, and

$H_{S,i}^-$ represents the i th negative sample in the negative sample queue.

[0054] In some embodiments, the first contrastive loss value $L_{NCE}^{V \rightarrow S}(H_V^m, H_S^+, K_S^-)$ is calculated according to Formula (2):

$$L_{NCE}^{V \rightarrow S}(H_V^m, H_S^+, K_S^-) = -\log \frac{\exp(\langle H_V^m, H_S^+ \rangle / t)}{\exp\left(\frac{\langle H_V^m, H_S^+ \rangle}{t}\right) + \sum_{i=1}^K \exp(\langle H_V^m, H_{S,i}^- \rangle / t)} \quad (2)$$

[0055] In the Formula (2), t is a hyper-parameter for controlling scaling. The operator $\langle A, B \rangle$ represents a cosine similarity of the vectors A and B .

[0056] In step 202, a second contrastive loss value is determined by using the first video feature, the second word segmentation feature and a preset second negative sample feature.

[0057] In some embodiments, the first video feature is converted to a global second positive sample feature H_V^+ by using the MLP model, and the second word segmentation feature is converted to a global text query feature H_S^m by using the MLP model. The second contrastive loss value is determined by using the text query feature H_S^m , the second positive sample feature H_V^+ , and a second negative sample feature K_V^- .

[0058] It should be noted that the second negative sample feature K_V^- is represented by the following Formula:

$$K_V^- = \{H_{V,i}^-\}_{i=1}^K \quad (3)$$

[0059] In the Formula (3), K represents a size of a negative sample queue included in the second negative sample feature and $H_{V,i}^-$ represents the i th negative sample in the negative sample queue.

[0060] In some embodiments, the second contrastive loss value $L_{NCE}^{S \rightarrow V}(H_S^m, H_V^+, K_V^-)$ is calculated according to Formula (4):

$$L_{NCE}^{S \rightarrow V}(H_S^m, H_V^+, K_V^-) = -\log \frac{\exp(\langle H_S^m, H_V^+ \rangle / t)}{\exp\left(\frac{\langle H_S^m, H_V^+ \rangle}{t}\right) + \sum_{i=1}^K \exp(\langle H_S^m, H_{V,i}^- \rangle / t)} \quad (4)$$

[0061] In the Formula (4), t is a hyper-parameter for controlling scaling. The operator $\langle A, B \rangle$ represents a cosine similarity of the vectors A and B .

[0062] In step 203, a first target is determined based on the first contrastive loss value and the second contrastive loss value.

[0063] In some embodiments, the first target is a sum of the first contrastive loss value and the second contrastive loss value. For example, the first target is calculated according to Formula (5). The first target is used to represent a combination of video-to-text and text-to-video video matching losses.

$$L_{Co-IM} = L_{NCE}^{V \rightarrow S}(H_V^m, H_S^+, K_S^-) + L_{NCE}^{S \rightarrow V}(H_S^m, H_V^+, K_V^-) \quad (5)$$

[0064] In step 204, a third contrastive loss value is determined by using the first video feature, the second video feature and the second negative sample feature.

[0065] In some embodiments, the third contrastive loss value is determined by using the video query feature H_V^m , the second positive sample feature H_V^+ , and the second negative sample feature K_V^- .

[0066] In some embodiments, the third contrastive loss value $L_{NCE}^V(H_V^m, H_V^+, K_V^-)$ is calculated according to Formula (6):

$$L_{NCE}^V(H_V^m, H_V^+, K_V^-) = -\log \frac{\exp(\langle H_V^m, H_V^+ \rangle / t)}{\exp\left(\frac{\langle H_V^m, H_V^+ \rangle}{t}\right) + \sum_{i=1}^K \exp(\langle H_V^m, H_{V,i}^- \rangle / t)} \quad (6)$$

[0067] In Formula (6), t is a hyper-parameter for controlling scaling. The operator $\langle A, B \rangle$ represents a cosine similarity of the vectors A and B .

[0068] In step 205, a fourth contrastive loss value is determined by using the first word segmentation feature, the second word segmentation feature, and the first negative sample feature.

[0069] In some embodiments, the fourth contrastive loss value

is determined by using the text query feature H_S^m , the first positive sample feature H_S^+ , and the first negative sample feature K_S^- .

[0070] In some embodiments, a fourth contrastive loss value $L_{NCE}^S(H_S^m, H_S^+, K_S^-)$ is calculated according to Formula (7):

$$L_{NCE}^S(H_S^m, H_S^+, K_S^-) = -\log \frac{\exp(\langle H_S^m, H_S^+ \rangle / t)}{\exp(\langle H_S^m, H_S^+ \rangle / t) + \sum_{i=1}^K \exp(\langle H_S^m, H_{S,i}^- \rangle / t)} \quad (7)$$

[0071] In Formula (7), t is a hyper-parameter for controlling scaling. The operator $\langle A, B \rangle$ represents a cosine similarity of the vectors A and B .

[0072] In step 206, a second target is determined based on the third contrastive loss value and the fourth contrastive loss value.

[0073] In some embodiments, the second target is a sum of the third contrastive loss value and the fourth contrastive loss value. The second target is calculated, for example, according to Formula (8). The second target is used to represent denoising losses within a video modality and within a text modality.

$$L_{Co-ID} = L_{NCE}^V(H_V^m, H_V^+, K_V^-) + L_{NCE}^S(H_S^m, H_S^+, K_S^-) \quad (8)$$

[0074] In step 207, a target function is determined based on the first target and the second target.

[0075] In some embodiments, the target function is a sum of the first target and the second target. For example, the target function L is calculated according to Formula (9):

$$L = L_{Co-IM} + L_{Co-ID} \quad (9)$$

[0076] Returning to FIG. 1. In step 106, a multi-modal pre-training is performed by using the pre-trained target function.

[0077] In the multi-modal pre-training method provided by the embodiment of the present disclosure, the pre-trained target function is determined based on a cross-modal matching loss and an intra-modal denoising loss, such that a correlation between

cross-modal data can be enhanced, and a comprehension capability of the multi-modal pre-training model to the contents of multi-modal data is effectively improved.

[0078] In some embodiments, the second video feature and the second word segmentation feature are fused to obtain a fused feature. The fused feature is inputted into an MLM (Masked Language Modelling) model to obtain a third target L_{MLM} , and the fused feature is inputted into an MSG (Masked Language Generation) model to obtain a fourth target L_{MSG} .

[0079] In some embodiments, the second video feature and the second word segmentation feature are fused by using a Cross-Modal Decoder to obtain a fused feature. The cross-modal decoder is used for outputting fused feature of the multi-modal information of the video and the text and providing feature inputs for subsequent tasks.

[0080] Since the cross-modal decoder is not an inventive focus of the present disclosure, it is not described here in detail.

[0081] In some embodiments, the target function L is determined based on a first target L_{Co-IM} , a second target L_{Co-ID} , a third target L_{MLM} , and a fourth target L_{MSG} .

[0082] In some embodiments, the target function L is a sum of the first target L_{Co-IM} , a second target L_{Co-ID} , a third target L_{MLM} , and a fourth target L_{MSG} .

[0083] For example, the target function L is calculated according to the following Formula (10):

$$L = L_{Co-IM} + L_{Co-ID} + L_{MLM} + L_{MSG} \quad (10)$$

[0084] FIG. 3 is a schematic structural diagram of a multi-modal pre-training apparatus according to one embodiment of the present disclosure. As shown in FIG. 3, the multi-modal pre-training apparatus comprises a first processing module 31, a second processing module 32, a third processing module 33, a

fourth processing module 34, a fifth processing module 35, and a sixth processing module 36.

[0085] The first processing module 31 is configured to sample the video in the video-text pair to obtain a first video frame sequence and is further configured to perform a word segmentation process on the text in the video-text pair to obtain a first word segmentation sequence.

[0086] In some embodiments, the video is sampled in an equidistant sampling manner to obtain the first video frame sequence.

[0087] In some embodiments, flags [CLS] and [SEP] are respectively provided at the beginning and end of the first word segmentation sequence, for convenience of subsequent processing.

[0088] The second processing module 32 is configured to mask the first video frame sequence to obtain a second video frame sequence and is further configured to mask the first word segmentation sequence to obtain a second word segmentation sequence.

[0089] In some embodiments, the video frames in the first video frame sequence are replaced with masks at random probabilities to obtain a second video frame sequence.

[0090] In some embodiments, the word segments in the first word segmentation sequence are replaced with masks at random probabilities to obtain a second word segmentation sequence.

[0091] The third processing module 33 is configured to encode the first video frame sequence to obtain a first video feature and is further configured to encode the first word segmentation sequence to obtain a first word segmentation feature.

[0092] In some embodiments, a video key-value encoder is used to encode the first video frame sequence to obtain a first video feature, and a sentence key-value encoder is used to encode the first word segmentation sequence to obtain a first word

segmentation feature.

[0093] The first video feature outputted by the video key-value encoder reflects contextual characteristics of unmasked video frames. The first word segmentation feature outputted by the sentence key-value encoder reflects contextual characteristics of the unmasked word segmentation sequence.

[0094] The fourth processing module 34 is configured to encode the second video frame sequence to obtain a second video feature and is further configured to encode the second word segmentation sequence to obtain a second word segmentation feature.

[0095] In some embodiments, a video query encoder is used to encode the second video frame sequence to obtain the second video features, and a sentence query encoder is used to encode the second word segmentation sequence to obtain the second word segmentation feature.

[0096] The second video feature outputted by the video query encoder reflects a correlation between frames in the video modality, and the second word segmentation feature outputted by the sentence query encoder reflects a correlation between words in the text modality.

[0097] The fifth processing module 35 is configured to determine a pre-trained target function by using the first video feature, the first word segmentation feature, the second video feature, and the second word segmentation feature. In some embodiments, the fifth processing module 35 determines a first contrastive loss value by using the first word segmentation feature, the second video feature, and a preset first negative sample feature.

[0098] For example, an MLP model is used to convert the first word segmentation feature into a global first positive sample feature H_s^+ . The MLP model is used to convert the second video feature into a global video query feature H_v^m . A first

contrastive loss value is determined by using the video query feature H_V^m , the first positive sample feature H_S^+ , and the first negative sample feature K_S^- .

[0099] In some embodiments, the first negative sample feature K_S^- is as shown in Formula (1) above.

[00100] In some embodiments, the first contrastive loss value $L_{NCE}^{V \rightarrow S}(H_V^m, H_S^+, K_S^-)$ is calculated by using Formula (2) above.

[00101] The fifth processing module 35 determines a second contrastive loss value by using the first video feature, the second word segmentation feature and a preset second negative sample feature. For example, the MLP model is used to convert the first video feature into a global second positive sample feature H_V^+ , and the MLP model is used to convert the second word segmentation feature into a global text query feature H_S^m . A second contrastive loss value is determined by using the text query feature H_S^m , the second positive sample feature H_V^+ , and the second negative sample feature K_V^- .

[00102] In some embodiments, the second negative sample feature K_V^- is as shown in Formula (3) above.

[00103] In some embodiments, the second contrastive loss value $L_{NCE}^{S \rightarrow V}(H_S^m, H_V^+, K_V^-)$ is calculated by using Formula (4) above.

[00104] The fifth processing module 35 determines a first target based on the first contrastive loss value and second contrastive loss value. In some embodiments, the first target is a sum of the first contrastive loss value and the second contrastive loss value. For example, the first target is calculated by using the above Formula (5). The first target is used to represent a combination of video-to-text and text-to-video matching losses.

[00105] The fifth processing module 35 determines a third contrastive loss value by using the first video feature, the second video feature, and the second negative sample feature.

In some embodiments, the third contrastive loss value is determined by using the video query feature H_V^m , the second positive sample feature H_V^+ , and the second negative sample feature K_V^- . For example, the third contrastive loss value $L_{NCE}^V(H_V^m, H_V^+, K_V^-)$ is calculated by using Formula (6) above.

[00106] The fifth processing module 35 determines a fourth contrastive loss value by using the first word segmentation feature, the second word segmentation feature and the first negative sample feature. In some embodiments, the fourth contrastive loss value is determined by using the text query feature H_S^m , the first positive sample feature H_S^+ , and the first negative sample feature K_S^- .

[00107] In some embodiments, the fourth contrastive loss value $L_{NCE}^S(H_S^m, H_S^+, K_S^-)$ is calculated by using Formula (7) above.

[00108] The fifth processing module 35 determines a second target based on the third contrastive loss value and the fourth contrastive loss value. In some embodiments, the second target is a sum of the third contrastive loss value and the fourth contrastive loss value. The second target is calculated, for example, by using the above Formula (8). The second target is used to represent denoising losses within a video modality and within a text modality.

[00109] The fifth processing module 35 determines a target function based on the first target and the second target. In some embodiments, the target function is a sum of the first target and the second target. For example, the target function L is calculated by using the above Formula (9).

[00110] In some embodiments, the fifth processing module 35 fuses the second video feature and the second word segmentation feature to obtain a fused feature. The fused feature is inputted to an MLM model to obtain a third target L_{MLM} , and the fused

feature is inputted to an MSG model to obtain a fourth target L_{MSG} .

[00111] In some embodiments, the second video feature and the second word segmentation feature are fused by using a cross-modal decoder to obtain a fused feature. The cross-modal decoder is used for outputting the fused feature of the video and text multi-modal information and providing characteristic input for subsequent tasks.

[00112] In some embodiments, the target function L is determined from the first target L_{Co-IM} , the second target L_{Co-ID} , the third target L_{MLM} and the fourth target L_{MSG} . In some embodiments, the target function L is a sum of the first target L_{Co-IM} , the second target L_{Co-ID} , the third target L_{MLM} and the fourth target L_{MSG} . For example, the target function L is calculated by using the above Formula (10).

[00113] The sixth processing module 36 is configured to perform multi-modal pre-training by using a pre-trained target function.

[00114] FIG. 4 is a schematic structural diagram of a multi-modal pre-training apparatus according to another embodiment of the present disclosure. As shown in FIG. 4, the multi-modal pre-training apparatus comprises a memory 41 and a processor 42.

[00115] The memory 41 is used for storing instructions, the processor 42 is coupled to the memory 41, and the processor 42 is configured to perform the method according to any one of the embodiments in FIG. 1 or FIG. 2 based on the instructions stored in the memory.

[00116] As shown in FIG. 4, the multi-modal pre-training apparatus further comprises a communication interface 43 for information interaction with other apparatuses. Meanwhile, the multi-modal pre-training apparatus further comprises a bus 44, wherein the processor 42, the communication interface 43 and

the memory 41 communicate with each other via the bus 44.

[00117] The memory 41 may comprise a high-speed RAM memory, and may also comprise a non-volatile memory, such as at least one disk memory. The memory 41 may also be a memory array. The memory 41 may also be partitioned into blocks, and the blocks may be combined into virtual volumes according to certain rules.

[00118] Further, the processor 42 may be a central processing unit CPU, or may be an application specific integrated circuit ASIC, or one or more integrated circuits configured to implement embodiments of the present disclosure.

[00119] The present disclosure also relates to a computer-readable storage medium, wherein the computer-readable storage medium stores computer instructions, which, when executed by a processor, implement the method according to any one of the embodiments in FIG. 1 or FIG. 2.

[00120] FIG. 5 is a schematic diagram of a multi-modal pre-training model according to one embodiment of the present disclosure.

[00121] As shown in FIG.5, a first video frame subsequence is obtained by sampling the video in the video-text pair, and a first word segmentation sequence is obtained by sampling the text in the video-text pair. The video frames in the first video frame sequence are replaced with masks at random probabilities to obtain a second video frame sequence. The word segments in the first word segmentation sequence are replaced with masks at random probabilities to obtain a second word segmentation sequence.

[00122] The first video frame sequence is encoded by using a video key-value encoder to obtain a first video feature, and the first word segmentation sequence is encoded by using a sentence key-value encoder to obtain a first word segmentation feature.

[00123] The second video frame sequence is encoded by using a video key-value encoder to obtain a second video feature, and the second word segmentation sequence is encoded by using a sentence key-value encoder to obtain a second word segmentation feature.

[00124] An MLP model is used for converting the first segmentation feature into a global first positive sample feature H_S^+ , the MLP model is used for converting the first video feature into a global second positive sample feature H_V^+ , the MLP model is used for converting the second video feature into a global video query feature H_V^m , and the MLP model is used for converting the second word segmentation feature into a global text query feature H_S^m . In a Co-IM (Contrastive Inter-modal Matching) module, a first contrastive loss value $L_{NCE}^{V \rightarrow S}(H_V^m, H_S^+, K_S^-)$ is determined by using the video query feature H_V^m , the first positive sample feature H_S^+ , and the first negative sample feature K_S^- according to the above Formula (2).

[00125] In some embodiments, the first negative sample feature K_S^- is as shown in Formula (1) above.

[00126] A second contrastive loss value $L_{NCE}^{S \rightarrow V}(H_S^m, H_V^+, K_V^-)$ is determined by using the text query feature H_S^m , the second positive sample feature H_V^+ and the second negative sample feature K_V^- according to the Formula (4).

[00127] In some embodiments, the second negative sample feature K_V^- is as shown in Formula (3) above.

[00128] Next, a first target L_{Co-IM} is calculated by using the above Formula (5).

[00129] In a Co-ID (Contrastive Intra-mode Denoising) module, a third contrast loss value $L_{NCE}^V(H_V^m, H_V^+, K_V^-)$ is determined by using the video query feature H_V^m , the second positive sample feature

H_V^+ , and the second negative sample feature K_V^- according to the above Formula (6).

[00130] A fourth contrastive loss value $L_{NCE}^S(H_S^m, H_S^+, K_S^-)$ is determined by using the text query feature H_S^m , the first positive sample feature H_S^+ , and the first negative sample feature K_S^- according to the Formula (7).

[00131] Next, a second target L_{Co-ID} is determined based on the third contrastive loss value and the fourth contrastive loss value according to the above Formula (8).

[00132] In addition, the second video feature and the second word segmentation feature are fused by using a cross-modal decoder to obtain a fused feature. The fused feature is inputted into an MLM model to obtain a third target L_{MLM} , and the fused feature is inputted into an MSG model to obtain a fourth target L_{MSG} .

[00133] Next, by using the above Formula (10), a target function L is obtained by taking a sum of the first target L_{Co-IM} , the second target L_{Co-ID} , the third target L_{MLM} , and the fourth target L_{MSG} .

[00134] In some embodiments, the functional unit modules described above can be implemented as a general purpose processor, a Programmable Logic Controller (PLC), a Digital Signal Processor (DSP), an Application Specific Integrated Circuit (ASIC), a Field-Programmable Gate Array (FPGA) or other programmable logic devices, discrete gate or transistor logics, discrete hardware components, or any suitable combinations thereof for performing the functions described in this disclosure.

[00135] It will be appreciated by those skilled in the art that all or part of the steps for implementing the above embodiments may be implemented by hardware, or may be implemented by a program instructing relevant hardware, where the program may be

stored in a computer-readable storage medium, and the storage medium may be a read-only memory, a magnetic disk or an optical disk.

[00136] The description of the present disclosure has been presented for purposes of illustration and description, and is not intended to be exhaustive or limited to the disclosure in the form disclosed. Many modifications and variations will be apparent to those skilled in this art. The embodiment was chosen and described in order to best explain principles and practical applications of the present disclosure, and to enable those skilled in the art to understand the disclosure so as to design various embodiments with various modifications as are suited to the particular use.

What is claimed is:

1. A multi-modal pre-training method, comprising:
 - sampling a video in a video-text pair to obtain a first video frame sequence;
 - performing word segmentation processing on a text in the video-text pair to obtain a first word segmentation sequence;
 - masking on the first video frame sequence to obtain a second video frame sequence;
 - masking on the first word segmentation sequence to obtain a second word segmentation sequence;
 - encoding the first video frame sequence to obtain a first video feature, and encoding the first word segmentation sequence to obtain a first word segmentation feature;
 - encoding the second video frame sequence to obtain a second video feature, and encoding the second word segmentation sequence to obtain a second word segmentation feature;
 - determining a pre-trained target function by using the first video feature, the first word segmentation feature, the second video feature and the second word segmentation feature; and
 - performing multi-modal pre-training by using the pre-trained target function.

2. The multi-modal pre-training method of claim 1, wherein the determining the pre-trained target function comprises:

- determining a first contrastive loss value by using the first word segmentation feature, the second video feature and a preset first negative sample feature;

- determining a second contrastive loss value by using the first video feature, the second word segmentation feature and a preset second negative sample feature;

- determining a first target based on the first contrastive

loss value and second contrastive loss value;

determining a third contrastive loss value by using the first video feature, the second video feature, and the second negative sample feature;

determining a fourth contrastive loss value by using the first word segmentation feature, the second word segmentation feature, and the first negative sample feature;

determining a second target according to the third contrastive loss value and the fourth contrastive loss value;
and

determining the target function according to the first target and the second target.

3. The multi-modal pre-training method of claim 2, wherein the determining the first contrastive loss value comprises:

converting the first word segmentation feature into a global first positive sample feature;

converting the second video feature into a global video query feature; and

determining the first contrastive loss value by using the video query feature, the first positive sample feature, and the first negative sample feature.

4. The multi-modal pre-training method of claim 3, wherein the determining the second contrastive loss value comprises:

converting the first video feature into a global second positive sample feature;

converting the second word segmentation feature into a global text query feature; and

determining the second contrastive loss value by using the text query feature, the second positive sample feature, and the second negative sample feature.

5. The multi-modal pre-training method of claim 4, wherein the determining the third contrastive loss value comprises:

determining the third contrastive loss value by using the video query feature, the second positive sample feature, and the second negative sample feature.

6. The multi-modal pre-training method of claim 5, wherein the determining the fourth contrastive loss value comprises:

determining the fourth contrastive loss value by using the text query feature, the first positive sample feature, and the first negative sample feature.

7. The multi-modal pre-training method of claim 2, wherein: the first target is a sum of the first contrastive loss value and the second contrastive loss value; and

the second target is a sum of the third contrastive loss value and the fourth contrastive loss value.

8. The multi-modal pre-training method of any one of claims 2-7, wherein

the target function is a sum of the first target and the second target.

9. The multi-modal pre-training method of any one of claims 2-7, further comprising:

fusing the second video feature and the second word segmentation feature to obtain a fused feature; and

inputting the fused feature into a masked language modelling MLM model to obtain a third target, and inputting the fused feature into a masked language generating MSG model to obtain a fourth target,

wherein the determining the target function according to the first target and the second target comprising:

determining the target function according to the first target, the second target, the third target and the fourth target.

10. The multi-modal pre-training method of claim 9, wherein the target function is a sum of the first target, the second target, the third target, and the fourth target.

11. A multi-modal pre-training apparatus, comprising:

a first processing module configured to sample a video in a video-text pair to obtain a first video frame sequence, and perform word segmentation processing on a text in the video-text pair to obtain a first word segmentation sequence;

a second processing module configured to mask on the first video frame sequence to obtain a second video frame sequence, and mask on the first word segmentation sequence to obtain a second word segmentation sequence;

a third processing module configured to encode the first video frame sequence to obtain a first video feature, and encode the first word segmentation sequence to obtain a first word segmentation feature;

a fourth processing module configured to encode the second video frame sequence to obtain a second video feature, and encode the second word segmentation sequence to obtain a second word segmentation feature;

a fifth processing module configured to determine a pre-trained target function by using the first video feature, the first word segmentation feature, the second video feature and the second word segmentation feature; and

a sixth processing module configured to perform multi-modal pre-training by using the pre-trained target function.

12. A multi-modal pre-training apparatus, comprising:

a memory; and

a processor coupled to the memory, which is configured to execute a multi-modal pre-training method according to any of claims 1-10 on a basis of instructions stored in the memory.

13. A non-transitory computer-readable storage medium, which stores a computer program that, when executed by a processor, performs a multi-modal pre-training method of any of claims 1-10.