

CROSS-MODAL INFORMATION RETRIEVAL METHOD AND DEVICE, AND STORAGE MEDIUM

CROSS-REFERENCE TO RELATED APPLICATION

[0001] The present application claims priority to Chinese Patent Application No. 201910099972.3, filed with the Chinese Patent Office on January 31, 2019 and entitled "Cross-Modal Information Retrieval Method and Device, and Storage Medium", the content of which is hereby incorporated by reference in its entirety.

TECHNICAL FIELD

[0002] The disclosure relates to the technical field of computers, and particularly to a cross-modal information retrieval method and device, and a storage medium.

BACKGROUND

[0003] Along with the development of computer networks, a user may acquire a large amount of information from a network. Due to a large amount of information, the user may retrieve information of interest by inputting a text or a picture. Along with constant optimization of information retrieval technology, a cross-modal retrieval manner has emerged. In the cross-modal retrieval manner, certain modality information may be used to search for other modality information with similar semantics. For example, a text corresponding to an image may be retrieved using the image. Alternatively, an image corresponding to a text may be retrieved using the text.

SUMMARY

[0004] In view of this, the present disclosure discloses technical solutions of cross-

modal information retrieval.

- [0005] According to an aspect of the disclosure, a cross-modal information retrieval method is provided, which may include the following operations.
- [0006] First modal information and second modal information are acquired.
- [0007] Feature fusion is performed on a modal feature of the first modal information and a modal feature of the second modal information to determine a first fused feature corresponding to the first modal information and a second fused feature corresponding to the second modal information.
- [0008] A similarity between the first modal information and the second modal information is determined based on the first fused feature and the second fused feature.
- [0009] In a possible implementation mode, the operation that feature fusion is performed on the modal feature of the first modal information and the modal feature of the second modal information to determine the first fused feature corresponding to the first modal information and the second fused feature corresponding to the second modal information may include the following operations.
- [0010] A fusion threshold parameter for feature fusion of the first modal information and the second modal information is determined based on the modal feature of the first modal information and the modal feature of the second modal information.
- [0011] Feature fusion is performed on the modal feature of the first modal information and the modal feature of the second modal information according to the fusion threshold parameter to determine the first fused feature corresponding to the first modal information and the second fused feature corresponding to the second modal information. The fusion threshold parameter is configured for fused features obtained by feature fusion according to a matching degree between features, and the fusion threshold parameter becomes smaller as the matching degree between the features is lower.
- [0012] In a possible implementation mode, the operation that the fusion threshold parameter for feature fusion of the first modal information and the second modal

information is determined based on the modal feature of the first modal information and the modal feature of the second modal information may include the following operations.

[0013] A second attention feature attended by the first modal information to the second modal information is determined according to the modal feature of the first modal information and the modal feature of the second modal information.

[0014] A first fusion threshold parameter corresponding to the first modal information is determined according to the modal feature of the first modal information and the second attention feature.

[0015] In a possible implementation mode, the operation that the second attention attended by the first modal information to the second modal information is determined may include the following operations.

[0016] The first modal information includes at least one information unit and the second modal information includes at least one information unit.

[0017] A first modal feature of each of the at least one information unit of the first modal information is acquired.

[0018] A second modal feature of each of the at least one information unit of the second modal information is acquired.

[0019] An attention weight between each information unit of the first modal information and each information unit of the second modal information is determined according to the first modal feature and the second modal feature.

[0020] A second attention feature attended by each information unit of the first modal information to the second modal information is determined according to the attention weight and the second modal feature.

[0021] In a possible implementation mode, the operation that the fusion threshold parameter for feature fusion of the first modal information and the second modal information is determined based on the modal feature of the first modal information and the modal feature of the second modal information may include the following operations.

- [0022] A first attention feature attended by the second modal information to the first modal information is determined according to the modal feature of the first modal information and the modal feature of the second modal information.
- [0023] A second fusion threshold parameter corresponding to the second modal information is determined according to the modal feature of the second modal information and the first attention feature.
- [0024] In a possible implementation mode, the operation that the first attention feature attended by the second modal information to the first modal information is determined according to the modal feature of the first modal information and the modal feature of the second modal information may include the following operations.
- [0025] The first modal information includes at least one information unit and the second modal information includes at least one information unit.
- [0026] The first modal feature of each of the at least one information unit of the first modal information is acquired.
- [0027] The second modal feature of each of the at least one information unit of the second modal information is acquired.
- [0028] The attention weight between each information unit of the first modal information and each information unit of the second modal information is determined according to the first modal feature and the second modal feature.
- [0029] A first attention feature attended by each information unit of the second modal information to the first modal information is determined according to the attention weight and the first modal feature.
- [0030] In a possible implementation mode, the operation that the first fused feature corresponding to the first modal information is determined may include the following operations.
- [0031] A second attention feature attended by the first modal information to the second modal information is determined according to the modal feature of the first modal

information and the modal feature of the second modal information.

[0032] Feature fusion is performed on the modal feature of the first modal information and the second attention feature by using the fusion threshold parameter to determine the first fused feature corresponding to the first modal information.

[0033] In a possible implementation mode, the operation that feature fusion is performed on the modal feature of the first modal information and the second attention feature by use of the fusion threshold parameter to determine the first fused feature corresponding to the first modal information may include the following operations.

[0034] Feature fusion is performed on the modal feature of the first modal information and the second attention feature to obtain a first fusion result.

[0035] The first fusion result is processed based on the fusion threshold parameter to obtain a processed first fusion result.

[0036] The first fused feature corresponding to the first modal information is determined based on the processed first fusion result and the first modal feature.

[0037] In a possible implementation mode, the operation that the second fused feature corresponding to the second modal information is determined may include the following operations.

[0038] A first attention feature attended by the second modal information to the first modal information is determined according to the modal feature of the first modal information and the modal feature of the second modal information.

[0039] The second fused feature corresponding to the second modal information is determined according to the modal feature of the second modal information and the first attention feature.

[0040] In a possible implementation mode, the operation that the second fused feature corresponding to the second modal information is determined according to the modal feature of the second modal information and the first attention feature may include the following operations.

- [0041] Feature fusion is performed on the modal feature of the second modal information and the first attention feature to obtain a second fusion result.
- [0042] The second fusion result is processed by using the fusion threshold parameter to obtain a processed second fusion result.
- [0043] The second fused feature corresponding to the second modal information is determined based on the processed second fusion result and the second modal feature.
- [0044] In a possible implementation mode, the operation that the similarity between the first modal information and the second modal information is determined based on the first fused feature and the second fused feature may include the following operation.
- [0045] The similarity between the first modal information and the second modal information is determined based on first attention information of the first fused feature and second attention information of the second fused feature.
- [0046] In a possible implementation mode, the first modal information may be information to be retrieved of a first modality, and the second modal information may be pre-stored information of a second modality. The method may include the following operation.
- [0047] The second modal information is determined as a retrieval result of the first modal information in condition that the similarity meets a preset condition.
- [0048] In a possible implementation mode, the second modal information may include multiple pieces of second modal information, and the operation that the second modal information is determined as the retrieval result of the first modal information in condition that the similarity meets the preset condition may include the following operations.
- [0049] The multiple pieces of second modal information are sequenced according to a similarity between the first modal information and each of the multiple pieces second modal information to obtain a sequencing result.
- [0050] The second modal information that the similarity meets the preset condition is

determined according to the sequencing result.

[0051] The second modal information that the similarity meets the preset condition is determined as the retrieval result of the first modal information.

[0052] In a possible implementation mode, the preset condition may include any one of the following conditions.

[0053] The similarity is greater than a preset value; and a rank of the similarity sequenced from low to high is higher than a preset rank.

[0054] In a possible implementation mode, the first modal information may include one piece of modal information in text information or image information; and the second modal information may include the other piece of modal information in the text information or the image information.

[0055] In a possible implementation mode, the first modal information may be training sample information of the first modality, and the second modal information may be training sample information of the second modality. Each piece of training sample information of the first modality and each piece of training sample information of the second modality may form a training sample pair.

[0056] In a possible implementation mode, the method may further include the following operations.

[0057] The training sample pairs include a positive sample pair and a negative sample pair.

[0058] A similarity of each training sample pair is acquired.

[0059] Loss in a feature fusion process of the first modal information and the second modal information is determined according to a similarity of a positive sample pair with a highest matching degree of modal information in positive sample pairs and a similarity of a negative sample pair with a lowest matching degree in negative sample pairs.

[0060] A model parameter of a cross-modal information retrieval model adopted for

the feature fusion process of the first modal information and the second modal information is regulated according to the loss.

[0061] According to another aspect of the disclosure, a cross-modal information retrieval device is provided, which may include an acquisition module, a fusion module and a determination module.

[0062] The acquisition module may be configured to acquire first modal information and second modal information.

[0063] The fusion module may be configured to perform feature fusion on a modal feature of the first modal information and a modal feature of the second modal information to determine a first fused feature corresponding to the first modal information and a second fused feature corresponding to the second modal information.

[0064] The determination module may be configured to determine a similarity between the first modal information and the second modal information based on the first fused feature and the second fused feature.

[0065] In a possible implementation mode, the fusion module may include a determination submodule and a fusion submodule.

[0066] The determination submodule may be configured to determine a fusion threshold parameter for feature fusion of the module, first modal information and the second modal information based on the modal feature of the first modal information and the modal feature of the second modal information.

[0067] The fusion submodule may be configured to perform feature fusion on the modal feature of the first modal information and the modal feature of the second modal information based on the fusion threshold parameter to determine the first fused feature corresponding to the first modal information and the second fused feature corresponding to the second modal information. The fusion threshold parameter is configured for fused features obtained by feature fusion according to a matching degree between features, and the fusion threshold parameter becomes smaller as the matching degree between the features is lower.

- [0068] In a possible implementation mode, the determination submodule may include a second attention determination unit and a first threshold determination unit.
- [0069] The second attention determination unit may be configured to determine a second attention feature attended by the first modal information to the second modal information according to the modal feature of the first modal information and the modal feature of the second modal information.
- [0070] The first threshold determination unit may be configured to determine a first fusion threshold parameter corresponding to the first modal information according to the modal feature of the first modal information and the second attention feature.
- [0071] In a possible implementation mode, the first modal information may include at least one information unit, and the second modal information may include at least one information unit. The second attention determination unit may be specifically configured to:
- [0072] acquire a first modal feature of each of the at least one information unit of the first modal information;
- [0073] acquire a second modal feature of each of the at least one information unit of the second modal information;
- [0074] determine an attention weight between each information unit of the first modal information and each information unit of the second modal information according to the first modal feature and the second modal feature; and
- [0075] determine a second attention feature attended by each information unit of the first modal information to the second modal information according to the attention weight and the second modal feature.
- [0076] In a possible implementation mode, the determination submodule may include a first attention determination unit and a second threshold determination unit.
- [0077] The first attention determination unit may be configured to determine a first attention feature attended by the second modal information to the first modal information

according to the modal feature of the first modal information and the modal feature of the second modal information.

[0078] The second threshold determination unit may be configured to determine a second fusion threshold parameter corresponding to the second modal information according to the modal feature of the second modal information and the first attention feature.

[0079] In a possible implementation mode, the first modal information may include at least one information unit, and the second modal information may include at least one information unit. The first attention determination unit may be specifically configured to:

[0080] acquire the first modal feature of each of the at least one information unit of the first modal information,

[0081] acquire the second modal feature of each of the at least one information unit of the second modal information,

[0082] determine an attention weight between each information unit of the first modal information and each information unit of the second modal information according to the first modal feature and the second modal feature and

[0083] determine a first attention feature attended by each information unit of the second modal information to the first modal information according to the attention weight and the first modal feature.

[0084] In a possible implementation mode, the fusion submodule may include a second attention determination unit and a first fusion unit.

[0085] The second attention determination unit may be configured to determine the second attention feature attended by the first modal information to the second modal information according to the modal feature of the first modal information and the modal feature of the second modal information.

[0086] The first fusion unit may be configured to perform feature fusion on the modal feature of the first modal information and the second attention feature by using the fusion

threshold parameter to determine the first fused feature corresponding to the first modal information.

[0087] In a possible implementation mode, the first fusion unit may be specifically configured to:

[0088] perform feature fusion on the modal feature of the first modal information and the second attention feature to obtain a first fusion result;

[0089] process, by using the fusion threshold parameter, the first fusion result to obtain a processed first fusion result; and

[0090] determine the first fused feature corresponding to the first modal information based on the processed first fusion result and the first modal feature.

[0091] In a possible implementation mode, the fusion submodule may include a first attention determination unit and a second fusion unit.

[0092] The first attention determination unit may be configured to determine the first attention feature of attended by the second modal information to the first modal information according to the modal feature of the first modal information and the modal feature of the second modal information.

[0093] The second fusion unit may be configured to determine the second fused feature corresponding to the second modal information according to the modal feature of the second modal information and the first attention feature.

[0094] In a possible implementation mode, the second fusion unit may be specifically configured to:

[0095] perform feature fusion on the modal feature of the second modal information and the first attention feature to obtain a second fusion result;

[0096] process the second fusion result by using the fusion threshold parameter to obtain a processed second fusion result; and

[0097] determine the second fused feature corresponding to the second modal

information based on the processed second fusion result and the second modal feature.

[0098] In a possible implementation mode, the determination module may be specifically configured to:

[0099] determine the similarity between the first modal information and the second modal information based on first attention information of the first fused feature and second attention information of the second fused feature.

[00100] In a possible implementation mode, the first modal information may be information to be retrieved of a first modality, and the second modal information may be pre-stored information of a second modality. The device may further include a retrieval result determination module.

[00101] The retrieval result determination module may be configured to determine the second modal information as a retrieval result of the first modal information in condition that the similarity meets a preset condition.

[00102] In a possible implementation mode, the second modal information may include multiple pieces of second modal information, and the retrieval result determination module may include a sequencing submodule, an information determination submodule and a retrieval result determination submodule.

[00103] The sequencing submodule may be configured to sequence the multiple pieces of second modal information according to a similarity between the first modal information and each of the multiple pieces of second modal information to obtain a sequencing result.

[00104] The information determination submodule may be configured to determine the second modal information that the similarity meets the preset condition according to the sequencing result.

[00105] The retrieval result determination submodule may be configured to determine the second modal information that the similarity meets the preset condition as the retrieval result of the first modal information.

- [00106] In a possible implementation mode, the preset condition may include any one of the following conditions.
- [00107] The similarity is greater than a preset value; and a rank of the similarity sequenced from low to high is higher than a preset rank.
- [00108] In a possible implementation mode, the first modal information may include one piece of modal information in text information or image information; and the second modal information may include the other piece of modal information in the text information or the image information.
- [00109] In a possible implementation mode, the first modal information may be training sample information of the first modality and the second modal information may be training sample information of the second modality. Each piece of training sample information of the first modality and each piece of training sample information of the second modality may form a training sample pair.
- [00110] In a possible implementation mode, the training sample pair may include a positive sample pair and a negative sample pair. The device may further include a feedback module, configured to:
- [00111] acquire a similarity of each training sample pair;
- [00112] determine a loss in a feature fusion process of the first modal information and the second modal information according to a similarity of a positive sample pair with a highest matching degree of modal information in positive sample pairs and a similarity of a negative sample pair with a lowest matching degree in negative sample pairs; and
- [00113] adjust, according to the loss, a model parameter of a cross-modal information retrieval model adopted for the feature fusion process of the first modal information and the second modal information.
- [00114] According to another aspect of the disclosure, a cross-modal information retrieval device is provided, which may include a processor and a memory configured to store instructions executable for the processor, where the processor is configured to

execute the abovementioned method.

[00115] According to another aspect of the disclosure, a non-transitory computer-readable storage medium is provided, in which computer program instructions may be stored, where the computer program instructions, when being executed by a processor, enable the process to implement the abovementioned method.

[00116] According to the embodiments of the disclosure, the first modal information and the second modal information are acquired, feature fusion is performed on the modal feature of the first modal information and the modal feature of the second modal information to determine the first fused feature corresponding to the first modal information and the second fused feature corresponding to the second modal information, and then the similarity between the first modal information and the second modal information is determined by use of the determined first fused feature and second fused feature. In this way, similarities between different modal information may be obtained by performing feature fusion on the different modal information. Compared with an existing technical solution that the similarities are determined by use of distances of features of the different modal information in the same vector space, in the embodiments of the disclosure, internal relations between the different modal information are considered, and the similarities between the different modal information are determined by performing feature fusion on the different modal information, so that the cross-modal information retrieval accuracy is improved.

[00117] Other features and aspects of the present disclosure will become apparent from the following detailed descriptions of exemplary embodiments with reference to the drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[00118] The accompanying drawings, which are incorporated in and constitute a part of this specification, illustrate embodiments consistent with the present disclosure and, together with the description, serve to explain the principles of the present disclosure.

- [00119] FIG. 1 is a flowchart of a cross-modal information retrieval method according to an embodiment of the present disclosure.
- [00120] FIG. 2 is a flowchart of determining fused features according to an embodiment of the present disclosure.
- [00121] FIG. 3 is a block diagram of image information including multiple image units according to an embodiment of the present disclosure.
- [00122] FIG. 4 is a block diagram of a process of determining a first attention feature according to an embodiment of the present disclosure.
- [00123] FIG. 5 is a block diagram of a process of determining a first fused feature according to an embodiment of the present disclosure.
- [00124] FIG. 6 is a flowchart of cross-modal information retrieval according to an embodiment of the present disclosure.
- [00125] FIG. 7 is a block diagram of training a cross-modal information retrieval model according to an embodiment of the present disclosure.
- [00126] FIG. 8 is a block diagram of a cross-modal information retrieval device according to an embodiment of the present disclosure.
- [00127] FIG. 9 is a block diagram of a cross-modal information retrieval device according to an exemplary embodiment of the present disclosure.

DETAILED DESCRIPTION

- [00128] Various exemplary embodiments, features and aspects of the disclosure will be described in detail below with reference to the drawings, in which the same reference numbers represent functionally the same or similar elements. Although various aspects of the embodiments are shown in the drawings, the drawings are not required to be drawn to scale unless otherwise specified.

[00129] In the embodiments of the disclosure, special term "exemplary" refers to "as an example, embodiment or description". Herein, any "exemplarily" described embodiment may not be interpreted to be superior to or better than other embodiments.

[00130] In addition, for describing the disclosure better, many specific details are presented in the following specific implementation modes. It should be understood that those skilled in the art may implement the disclosure even without some specific details. In some examples, methods, means, components and circuits which are well known to those skilled in the art are not described in detail, so as to highlight the subject matter of the disclosure.

[00131] The following method, device, electronic device or storage medium of the embodiments of the disclosure may be applied to any scenario requiring cross-modal information retrieval, and for example, may be applied to retrieval software and information positioning. A specific application scenario is not limited in the embodiments of the disclosure, and any solution for implementing cross-modal information retrieval by use of the method provided in the embodiments of the disclosure shall fall within the scope of protection of the disclosure.

[00132] According a cross-modal information retrieval solution provided in the embodiments of the disclosure, first modal information and second modal information may be acquired respectively, and then feature fusion may be performed on a modal feature of the first modal information and a modal feature of the second modal information based on the modal feature of the first modal information and the modal feature of the second modal information to obtain a first fused feature corresponding to the first modal information and a second fused feature corresponding to the second modal information, so that the correlation between the first modal information and the second modal information may be considered. In this way, when a similarity between the first modal information and the second modal information is determined, the similarity between different modal information may be evaluated by use of the obtained two fused features, and the correlations between the different modal information may be considered, so that the cross-modal information retrieval accuracy is improved.

[00133] In related art, during cross-modal information retrieval, a similarity between a

text and an image is usually determined according to feature vectors of the text and the image in the same vector space, which, however, does not take internal relation between different modal information into account. For example, nouns in the text may usually correspond to some regions in the image. For another example, quantifiers in the text may correspond to some specific objects in the image. It is apparent that an internal relation between cross-modal information is not considered in the present cross-modal information retrieval manner, resulting in inaccuracy of a cross-modal information retrieval result. In the embodiments of the disclosure, the internal relation between cross-modal information is considered, so that the accuracy of a cross-modal information retrieval process is improved. The cross-modal information retrieval solution provided in the embodiments of the disclosure will be described below in combination with the drawings in detail.

[00134] FIG. 1 is a flowchart of a cross-modal information retrieval method according to an embodiment of the disclosure. As shown in FIG. 1, the method includes the following steps.

[00135] In block 11, first modal information and second modal information are acquired.

[00136] In the embodiment of the disclosure, a retrieval device (for example, a retrieval device like retrieval software, a retrieval platform and a retrieval server) may acquire the first modal information or the second modal information. For example, the retrieval device acquires the first modal information or second modal information transmitted by user equipment. For another example, the retrieval device acquires the first modal information or the second modal information according to user operations. The retrieval platform may also acquire the first modal information or the second modal information from a local storage or a database. Herein, the first modal information and the second modal information are information of different modalities. For example, the first modal information may include one type of modal information in text information or image information and the second modal information may include one type of modal information in the text information or the image information. Herein, the first modal information and the second modal information are not limited to the image information

and the text information, and may also include voice information, video information and optical signal information, etc. Herein, the modality may be understood as a type or presentation form of the information. The first modal information and the second modal information may be information of different modalities.

[00137] In block 12, feature fusion is performed on a modal feature of the first modal information and a modal feature of the second modal information to determine a first fused feature corresponding to the first modal information and a second fused feature corresponding to the second modal information.

[00138] After the first modal information and the second modal information are acquired, feature extraction may be performed on the first modal information and the second modal information to determine the modal feature of the first modal information and the modal feature of the second modal information respectively. The modal feature of the first modal information may form a first modal feature vector, and the modal feature of the second modal information may form a second modal feature vector. Then, feature fusion may be performed on the first modal information and the second modal information according to the first modal feature vector and the second modal feature vector. When feature fusion is performed on the first modal information and the second modal information, the first modal feature vector and the second modal feature vector may be mapped to feature vectors in the same vector space at first, and then feature fusion is performed on the two feature vectors obtained by mapping. Such feature fusion manner is simple, but a matching degree between the features of the first modal information and the second modal information cannot be acquired well. The embodiment of the disclosure also provides another feature fusion manner to acquire the matching degree between the features of the first modal information and the second modal information well.

[00139] FIG. 2 is a flowchart of determining fused features according to an embodiment of the disclosure. The following steps may be included.

[00140] In block 121, a fusion threshold parameter for feature fusion of the platform, first modal information and the second modal information is determined based on the modal feature of the first modal information and the modal feature of the second modal

information.

[00141] In block 122, feature fusion is performed on the modal feature of the first modal information and the modal feature of the second modal information based on fusion threshold parameter to determine the first fused feature corresponding to the first modal information and the second fused feature corresponding to the second modal information. The fusion threshold parameter is configured for fused features obtained by feature fusion according to a matching degree between features, and the fusion threshold parameter becomes smaller as the matching degree between the features is lower.

[00142] Herein, when feature fusion is performed on the modal feature of the first modal information and the modal feature of the second modal information, the fusion threshold parameter for feature fusion of the modal feature of the first modal information and the modal feature of the second modal information may be determined at first according to the modal feature of the first modal information and the modal feature of the second modal information, and then feature fusion is performed on the first modal information and the second modal information by use of the fusion threshold parameter. The fusion threshold parameter may be set according to the matching degree between the features, where the feature fusion parameter is greater if the matching degree between the features is higher. Therefore, in a feature fusion process, matched features are reserved and mismatched features are filtered, and the first fused feature corresponding to the first modal information and the second fused feature corresponding to the second modal information are determined. Setting the fusion threshold parameter in the feature fusion process enables to acquire the matching degree between the features of the first modal information and the second modal information well in a cross-modal information retrieval process.

[00143] Given that the first modal information and the second modal information may be fused better based on the fusion threshold parameter, a process of determining the fusion threshold parameter will be described below.

[00144] In a possible implementation mode, the fusion threshold parameter may include a first fusion threshold parameter and a second fusion threshold parameter. The first fusion threshold parameter may correspond to the first modal information, and the

second fusion threshold parameter may correspond to the second modal information. When the fusion threshold parameter is determined, the first fusion threshold parameter and the second fusion threshold parameter may be determined respectively. When the first fusion threshold parameter is determined, a second attention feature attended by the first modal information to the second modal information may be determined according to the modal feature of the first modal information and the modal feature of the second modal information, and then the first fusion threshold parameter corresponding to the first modal information is determined according to the modal feature of the first modal information and the second attention feature. Correspondingly, when the second fusion threshold parameter is determined, a first attention feature attended by the second modal information to the first modal information may be determined according to the modal feature of the first modal information and the modal feature of the second modal information, and then the second fusion threshold parameter corresponding to the second modal information is determined according to the modal feature of the second modal information and the first attention feature.

[00145] Herein, the first modal information may include at least one information unit, and correspondingly, the second modal information may include at least one information unit. Each information unit may have the same or different size, and there may be an overlap between each information unit. For example, under the condition that the first modal information or the second modal information is image information, the image information may include multiple image units, each image unit may have the same or different size, and there may be an overlap between each image unit. FIG. 3 is a block diagram of image information including multiple image units according to an embodiment of the disclosure. As shown in FIG. 3, an image unit *a* corresponds to a hat region of a person, an image unit *b* corresponds to an ear region of the person, and an image unit *c* corresponds to an eye region of the person. The image units *a*, *b* and *c* have different sizes, and there is an overlapping part between the image unit *a* and the image unit *b*.

[00146] In a possible implementation mode, when determining the second attention feature attended by the first modal information to the second modal information, the retrieval device may acquire a first modal feature of each information unit of the first

modal information and acquire a second modal feature of each information unit of the second modal information. Then, an attention weight between each information unit of the first modal information and each information unit of the second modal information is determined according to the first modal feature and the second modal feature, and a second attention feature attended by each information unit of the first modal information to the second modal information is determined according to the attention weight and the second modal feature.

[00147] Correspondingly, when determining the first attention feature attended by the second modal information to the first modal information, the retrieval device may acquire the first modal feature of each information unit of the first modal information and acquire the second modal feature of each information unit of the second modal information. Then, the attention weight between each information unit of the first modal information and each information unit of the second modal information is determined according to the first modal feature and the second modal feature, and a first attention feature attended by each information unit of the second modal information to the first modal information is determined according to the attention weight and the first modal feature.

[00148] FIG. 4 is a block diagram of a process of determining a first attention feature according to an embodiment of the disclosure. For example, the first modal information is image information and the second modal information is text information. The retrieval device may acquire an image feature vector of each image unit of the image information (which is an example of the first modal feature). The image feature vector of the image unit may be represented as formula (1):

[00149]
$$V = [v_1, v_2, \dots, v_i, \dots, v_R] \in \mathbb{R}^{d \times R} \quad (1);$$

[00150] where R is the number of the image units, d is a dimension of the image feature vector, v_i is the image feature vector of the i-th image unit, and $\mathbb{R}$ represents a real matrix. Correspondingly, the retrieval device may acquire a text feature vector of each text unit of the text information (which is an example of the second modal feature). The text feature vector of the text unit may be represented as formula (2):

[00151]
$$S = [s_1, s_2, \dots, s_j, \dots, s_T] \in \mathbb{R}^{d \times T} \quad (2);$$

[00152] where T is the number of the text units, d is a dimension of the text feature vector, and s_j is the text feature vector of the j -th text unit. The retrieval device may determine an association matrix between the image feature vectors and the text feature vectors according to the image feature vectors and the text feature vectors, and then determine an attention weight between each image unit of the image information and each text unit of the text information by using the association matrix. MATMUL in FIG. 4 denotes a multiplication operation.

[00153] Herein, the association matrix may be represented as formula (3):

[00154]
$$A = (\tilde{W}_v V)^T (\tilde{W}_s S) \quad (3);$$

[00155] where $\tilde{W}_v, \tilde{W}_s \in \mathbb{R}^{d_h \times d}$, and d_h is a dimension of the matrix $\tilde{W}_v$ and $\tilde{W}_s$. $\tilde{W}_v$ is a mapping matrix for mapping an image feature to a d_h -dimensional vector space, and $\tilde{W}_s$ is a mapping matrix for mapping a text feature to the d_h -dimensional vector space.

[00156] The attention weight between the image unit and the text unit, that is determined by use of the association matrix, may be represented as formula (4):

[00157]
$$\tilde{A}_v = \text{softmax} \frac{A^T}{\sqrt{d_h}} \quad (4);$$

[00158] where the i -th row of $\tilde{A}_v$ represents an attention weight of the i -th text unit for an image unit, and softmax represents a normalization exponential function.

[00159] After the attention weight between the image unit and the text unit is obtained, a first attention feature attended by each text unit to the image information may be determined according to the attention weight and the image feature. The first attention feature attended by the text unit to the image information may be represented as formula (5):

[00160]
$$\tilde{\mathbf{V}} = \tilde{\mathbf{A}}_v \mathbf{V}^T \in \mathbb{R}^{T \times d} \quad (5);$$

[00161] where the i -th row of $\tilde{\mathbf{V}}$ represents an attention weight of the image feature attended by the i -th text unit, i being a positive integer less than or equal to T .

[00162] Correspondingly, the attention weight between the text unit and the image unit, that is determined by use of the association matrix, may be represented as $\tilde{\mathbf{A}}_s$. The first attention feature $\tilde{\mathbf{S}} \in \mathbb{R}^{R \times d}$ attended by the text unit to the image information may be obtained according to $\tilde{\mathbf{A}}_s$ and $\mathbf{S}$, where the j -th row of $\tilde{\mathbf{S}}$ may represent an attention weight of the text feature attended by the j -th image unit, j being a positive integer less than or equal to R .

[00163] In the embodiment of the disclosure, after determining the first attention feature and the second attention feature, the retrieval device may determine the first fusion threshold parameter corresponding to the first modal information according to the modal feature of the first modal information and the second attention feature, and determine the second fusion threshold parameter corresponding to the second modal information according to the modal feature of the second modal information and the first attention feature. A process of determining the first fusion threshold parameter and the second fusion threshold parameter will be described below.

[00164] For example, the first modal information is image information and the second modal information is text information. The first attention feature may be $\tilde{\mathbf{V}}$, and the second attention feature may be $\tilde{\mathbf{S}}$. When the first fusion threshold parameter corresponding to the image information is determined, it may be determined according to the following formula (6):

[00165]
$$\mathbf{g}_i = \sigma(\mathbf{v}_i \odot \tilde{\mathbf{s}}_i), \quad i \in \{1, \dots, R\} \quad (6);$$

[00166] where $\odot$ denotes the element-wise product, $\sigma(\cdot)$ denotes a sigmoid function, and $\mathbf{g}_i \in \mathbb{R}^{d \times 1}$ denotes the fusion threshold between $\mathbf{v}_i$ and $\tilde{\mathbf{s}}_i$. The fusion threshold is

greater if a matching degree between an image unit and the text information is higher, and thus the fusion operation may be facilitated. On the contrary, The fusion threshold value is smaller if a matching degree between an image unit and the text information is lower, and thus the fusion operation may be suppressed.

[00167] A first fusion threshold parameter corresponding to each image unit of the image information may be represented as formula (7):

$$[00168] \quad G_v = [g_1, \dots, g_R] \in \mathbb{R}^{d \times R} \quad (7);$$

[00169] In the same manner, a second fusion threshold parameter corresponding to each text unit of the text information may be obtained as formula (8):

$$[00170] \quad H_s = [h_1, \dots, h_R] \in \mathbb{R}^{d \times T} \quad (8);$$

[00171] In the embodiment of the disclosure, after determining the fusion threshold parameter, the retrieval device may perform feature fusion on the first modal information and the second modal information by use of the fusion threshold parameter. A process of feature fusion between the first modal information and the second modal information will be described below.

[00172] In a possible implementation mode, the second attention feature attended by the first modal information to the second modal information may be determined according to the modal feature of the first modal information and the modal feature of the second modal information. Then, feature fusion is performed on the modal feature of the first modal information and the second attention feature by use of the fusion threshold parameter to determine the first fusion threshold parameter corresponding to the first modal information.

[00173] Herein, during feature fusion, feature fusion may be performed on the modal feature of the first modal information and the second attention feature. In this way, attention information between the first modal information and the second modal information is considered, and an internal relation between the first modal information and the second modal information also is considered, so that feature fusion of the first

modal information and the second modal information is implemented better.

[00174] In a possible implementation mode, when feature fusion is performed on the modal feature of the first modal information and the second attention feature by use of the fusion threshold parameter to determine the first fused feature corresponding to the first modal information, feature fusion may first be performed on the modal feature of the first modal information and the second attention feature to obtain a first fusion result. Then, the fusion threshold parameter is applied to the first fusion result to obtain a processed first fusion result, and the first fused feature corresponding to the first modal information is determined based on the processed first fusion result and the first modal feature.

[00175] The fusion threshold parameter may include a first fusion threshold parameter and a second fusion threshold parameter. When feature fusion is performed on the modal feature of the first modal information and the second attention feature, the first fusion threshold parameter may be used, namely the first fusion threshold parameter may be caused to act on the first fusion result to determine the first fused feature.

[00176] A process of determining the first fused feature corresponding to the first modal information in the embodiment of the disclosure will be described below in combination with the drawings.

[00177] FIG. 5 is a block diagram of a process of determining a first fused feature according to an embodiment of the disclosure.

[00178] For example, the first modal information is the image information and the second modal information is the text information. The image feature vector (which is an example of the first modal feature) of each image unit of the image information is V , and a first attention feature vector formed by the first attention feature of the image information may be $\tilde{V}$. The text feature vector (which is an example of the second modal feature) of each text unit of the text information is S , and a second attention feature vector formed by the second attention feature of the image information may be $\tilde{S}$. The retrieval device may perform feature fusion on the image feature vector V and the second attention feature vector $\tilde{S}$ to obtain a first fusion result $V \oplus \tilde{S}$, then apply a first fusion

parameter G_v to $V \oplus \tilde{S}$ to obtain a processed first fusion result $G_v \odot V \oplus \tilde{S}$; and obtain the first fused feature according to the processed first fusion result $G_v \odot V \oplus \tilde{S}$ and the image feature vector V .

[00179] The first fused feature may be represented as formula (9):

$$[00180] \quad \bar{V} = \text{ReLU}(\hat{W}_v(G_v \odot (V \oplus \tilde{S})) + \hat{b}_v) + V \quad (9);$$

[00181] where $\hat{W}_v$ and $\hat{b}_v$ are fusion parameters corresponding to the image information, $\odot$ denotes element-wise product, $\oplus$ is the fusion operation, and ReLU denotes a linear rectification operation.

[00182] In a possible implementation mode, the first attention feature attended by the second modal information to the first modal information may be determined according to the modal feature of the first modal information and the modal feature of the second modal information. Then feature fusion is performed on the modal feature of the second modal information and the first attention feature by use of the fusion threshold parameter to determine the second fusion threshold parameter corresponding to the second modal information.

[00183] During feature fusion, feature fusion may be performed on the modal feature of the second modal information and the first attention feature. In this way, the attention information between the first modal information and the second modal information is considered, and the internal relation between the first modal information and the second modal information is also considered, so that feature fusion of the first modal information and the second modal information is implemented better.

[00184] Herein, when feature fusion is performed on the modal feature of the second modal information and the first attention feature by use of the fusion threshold parameter to determine the second fused feature corresponding to the second modal information, feature fusion is first performed on the modal feature of the second modal information and the first attention feature to obtain a second fusion result. Then, the second fusion result is processed by using the fusion threshold parameter to obtain a processed second

fusion result, and the second fused feature corresponding to the second modal information is determined based on the processed second fusion result and the second modal feature.

[00185] Herein, when feature fusion is performed on the modal feature of the first modal information and the second attention feature, the second fusion threshold parameter may be used, namely the second fusion threshold parameter may be applied to the second fusion result to determine the second fused feature.

[00186] The process of determining the second fused feature is similar to the process of determining the first fused feature and will not be elaborated herein. For example, the second modal information is the text information, and a second fused feature vector formed by the second fused feature may be represented as formula (10):

$$[00187] \quad \hat{S} = \text{ReLU}(\hat{W}_s(H_s \odot (S \oplus \tilde{V})) + \hat{b}_s) + S \quad (10);$$

[00188] where $\hat{W}_s$ and $\hat{b}_s$ are fusion parameters corresponding to the text information, $\odot$ denotes element-wise product, $\oplus$ denotes the fusion operation, and ReLU denotes the linear rectification operation.

[00189] In block 13, a similarity between the first modal information and the second modal information is determined based on the first fused feature and the second fused feature.

[00190] In the embodiment of the disclosure, the retrieval device may determine the similarity between the first modal information and the second modal information according to the first fused feature vector formed by the first fused feature and the second fused feature vector formed by the second fused feature. For example, feature fusion operation may be performed on the first fused feature vector and the second fused feature vector, or, a matching operation and the like may be performed on the first fused feature vector and the second fused feature vector, so as to determine the similarity between the first modal information and the second modal information. For obtaining a more accurate similarity, the embodiment of the disclosure also provides a manner for determining the similarity between the first modal information and the second modal information. A

process of determining the similarity in the embodiment of the disclosure will be described below.

[00191] In a possible implementation mode, when the similarity between the first modal information and the second modal information is determined, first attention information of the first fused feature may be acquired, and second attention information of the second fused feature may be acquired. Then, the similarity between the first modal information and the second modal information is determined based on the first attention information of the first fused feature and the second attention information of the second fused feature.

[00192] For example, under the condition that the first modal information is the image information, the first fused feature vector $\hat{V}$ of the image information corresponds to R image units. When the first attention information is determined according to the first fused feature vector, attention information of different image units may be extracted by use of multiple attention branches. For example, there are M attention branches, and a processing process of each attention branch is represented as formula (11):

$$[00193] \quad A_v^{*(i)} = softmax(\frac{W_v^{*(i)} \hat{V}}{\sqrt{d}}) \quad (11);$$

[00194] where $W_v^{*(i)}$ denotes a linear mapping parameter, $i \in \{1, \dots, M\}$ represents the i-th attention branch, $A_v^{*(i)}$ represents the attention information for R image units from the i-th attention branch, *softmax* represents a normalization exponential function, and $1/\sqrt{d}$ represents a weight parameter, which is capable of controlling a magnitude of the attention information to ensure that the obtained attention information is in a proper magnitude range.

[00195] Then, the attention information from each of the M attention branches may be aggregated, and the aggregated attention information is averaged to obtain final first attention information of the first fused feature.

[00196] The first attention information may be represented as formula (12):

[00197]
$$\hat{v} = \text{SAM}(\hat{V}) = \sum_{i=1}^M A_v^{*(i)} \hat{V}^T \quad (12).$$

[00198] Correspondingly, the second attention information may be $\hat{s}$.

[00199] The similarity between the first modal information and the second modal information may be represented as formula (13):

[00200]
$$m = \hat{s}^T \hat{v} \quad (13);$$

[00201] where m is within a range between 0 and 1, 1 represents that the first modal information and the second modal information are matched, and 0 represents that the first modal information and the second modal information are mismatched. The matching degree of the first modal information and the second modal information may be determined according to a distance between m and 0 or 1.

[00202] In the abovementioned cross-modal information retrieval manner, considering the internal relation between the different modal information, the similarity between the different modal information is determined by performing feature fusion on the different modal information, so that the cross-modal information retrieval accuracy is improved.

[00203] FIG. 6 is a flowchart of cross-modal information retrieval according to an embodiment of the disclosure. The first modal information may be information to be retrieved of a first modality, and the second modal information may be pre-stored information of a second modality. The cross-modal information retrieval method may include the following steps.

[00204] In block 61, first modal information and second modal information are acquired.

[00205] In block 62, feature fusion is performed on a modal feature of the first modal information and a modal feature of the second modal information to determine a first fused feature corresponding to the first modal information and a second fused feature corresponding to the second modal information.

[00206] In block 63, a similarity between the first modal information and the second

modal information is determined based on the first fused feature and the second fused feature.

[00207] In block 64, under the condition that the similarity meets a preset condition, the second modal information is determined as a retrieval result of the first modal information.

[00208] Herein, a retrieval device may acquire the first modal information input by a user and acquire the second modal information from a local storage or a database. Responsive to determining that the similarity between the first modal information and the second modal information meets the preset condition through the above steps, the second modal information may be determined as the retrieval result of the first modal information.

[00209] In a possible implementation mode, there are multiple pieces of second modal information. When the second modal information is determined as the retrieval result of the first modal information, the multiple pieces of second modal information may be sequenced according to a similarity between the first modal information and each piece of second modal information to obtain a sequencing result. The second modal information that the similarity meets the preset condition may be determined according to the sequencing result of the second modal information, and the second modal information that the similarity meets the preset condition is determined as the retrieval result of the first modal information.

[00210] The preset condition includes any one of the following conditions.

[00211] The similarity is greater than a preset value; and a rank of the similarity sequenced from low to high is higher than a preset rank.

[00212] For example, when the second modal information is determined as the retrieval result of the first modal information, if the similarity between the first modal information and second modal information, the second modal information is determined as the retrieval result of the first modal information. Or, when the second modal information is determined as the retrieval result of the first modal information, the multiple pieces of second modal information may be sequenced according to the similarity between the first

modal information and each piece of second modal information and according to the similarity sequence from large to small to obtain the sequencing result, and then the second modal information of which the rank is higher than the preset rank is determined as the retrieval result of the first modal information according to the sequencing result. For example, the second modal information with the highest rank is determined as the retrieval result of the first modal information, namely the second modal information corresponding to the highest similarity may be determined as the retrieval result of the first modal information. Herein, there may be one or more retrieval results.

[00213] After the second modal information is determined as the retrieval result of the first modal information, the retrieval result may be output to a user side. For example, the retrieval result may be sent to the user side, or, the retrieval result is displayed on a display interface.

[00214] FIG. 7 is a block diagram of a training process of a cross-modal information retrieval modal according to an embodiment of the disclosure. The first modal information may be training sample information of the first modality, the second modal information may be training sample information of the second modality, and each piece of the training sample information of the first modality and each piece of the training sample information of the second modality form a training sample pair.

[00215] In the training process, each training sample pair may be input to the cross-modal information retrieval model. For example, the training sample pair is an image-text pair. An image sample and text sample in the image-text pair may be input to the cross-modal information retrieval model respectively, and modal features of the image sample and modal features of the text sample are extracted by use of the cross-modal information retrieval model. Or, an image feature of the image sample and a text feature of the text sample are input to the cross-modal information retrieval model. Then, the first attention feature $\tilde{V}$ and second attention information $\tilde{S}$ co-attended by both the first modal information and the second modal may be determined by use of a cross-modal attention layer of the cross-modal information retrieval model, and feature fusion is performed on the first modal information and the second modal information by use of a threshold feature fusion layer to obtain the first fused feature $\hat{V}$ corresponding to the first

modal information and the second fused feature $\hat{S}$ corresponding to the second modal information. Next, the first attention information $\hat{V}$ self-attended by the first fused feature $\hat{V}$ and the second attention information $\hat{S}$ self-attended by the second fused feature $\hat{S}$ are determined by use of self-attention layer. Finally, the similarity m between the first modal information and the second modal information is output by using a Multi-Layer Perceptron (MLP) and sigmoid function (sigmoid σ).

[00216] Herein, the training sample pair may include a positive sample pair and a negative sample pair. In the process of training the cross-modal information retrieval model, loss of the cross-modal information retrieval model may be obtained by use of a loss function, so as to adjust a parameter of the cross-modal information retrieval model according to the obtained loss.

[00217] In a possible implementation mode, a similarity of each training sample pair may be acquired, then the loss in the feature fusion of the first modal information and the second modal information is determined according to the similarity of the positive sample pair with a highest modal information matching degree in the positive sample pairs and the similarity of the negative sample pair with a lowest matching degree in the negative sample pairs. The model parameters of the cross-modal information retrieval model adopted for the feature fusion of the first modal information and the second modal information are adjusted according to the loss. In the implementation mode, the loss in the training process is determined according to the similarity of the positive sample pair with the highest matching degree and the similarity of the negative sample pair with the lowest matching degree, so that the cross-modal information retrieval accuracy of the cross-modal information retrieval model is improved.

[00218] The loss of the cross-modal information retrieval model may be determined according to the following formula (14):

[00219]
$$\mathcal{L}_{BCE-h}(\mathcal{L}, \mathcal{C}) = \log(m(\mathcal{L}, \mathcal{C})) + \max_{\mathcal{C}'} [\log(1 - m(\mathcal{L}, \mathcal{C}'))] + \log(m(\mathcal{L}, \mathcal{C})) + \max_{\mathcal{L}'} [\log(1 - m(\mathcal{L}', \mathcal{C}))] \quad (14);$$

[00220] where $\mathcal{L}_{BCE-h}(\mathcal{L}, \mathcal{C})$ is the calculated loss, $m(\mathcal{L}, \mathcal{C})$ represents the similarity

between the sample pairs, $(\mathcal{L}, \mathcal{C})$ is a group of positive sample pairs, and $(\mathcal{L}, \mathcal{C}')$ and $(\mathcal{L}', \mathcal{C})$ are respective negative sample pairs.

[00221] Through the process of training the cross-modal information retrieval model, the loss in the training process is determined by use of the similarity of the positive sample pair with the highest matching degree and the similarity of the negative sample pair with the lowest matching degree, so that the accuracy that cross-modal information retrieval model retrieves the cross-modal information is improved.

[00222] FIG. 8 is a block diagram of a cross-modal information retrieval device according to an embodiment of the disclosure. As shown in FIG. 8, the cross-modal information retrieval device includes an acquisition module 81, a fusion module 82 and a determination module 83.

[00223] The acquisition module 81 is configured to acquire first modal information and second modal information.

[00224] The fusion module 82 is configured to perform feature fusion on a modal feature of the first modal information and a modal feature of the second modal information to determine a first fused feature corresponding to the first modal information and a second fused feature corresponding to the second modal information.

[00225] The determination module 83 is configured to determine a similarity between the first modal information and the second modal information based on the first fused feature and the second fused feature.

[00226] In a possible implementation mode, the fusion module 82 includes a determination submodule and a fusion submodule.

[00227] The determination submodule is configured to determine a fusion threshold parameter for feature fusion of the first modal information and the second modal information based on the modal feature of the first modal information and the modal feature of the second modal information.

[00228] The fusion submodule is configured to perform feature fusion on the modal

feature of the first modal information and the modal feature of the second modal information based on the fusion threshold parameter to determine the first fused feature corresponding to the first modal information and the second fused feature corresponding to the second modal information. The fusion threshold parameter is configured for fused features obtained by feature fusion according to a matching degree between features, and the fusion threshold parameter becomes smaller as the matching degree between the features is lower.

[00229] In a possible implementation mode, the determination submodule includes a second attention determination unit and a first threshold determination unit.

[00230] The second attention determination unit is configured to determine a second attention feature attended by the first modal information to the second modal information according to the modal feature of the first modal information and the modal feature of the second modal information.

[00231] The first threshold determination unit is configured to determine a first fusion threshold parameter corresponding to the first modal information according to the modal feature of the first modal information and the second attention feature.

[00232] In a possible implementation mode, the first modal information includes at least one information unit, and the second modal information includes at least one information unit. The second attention determination unit is specifically configured to:

[00233] acquire a first modal feature of each information unit of the first modal information,

[00234] acquire a second modal feature of each information unit of the second modal information,

[00235] determine an attention weight between each information unit of the first modal information and each information unit of the second modal information according to the first modal feature and the second modal feature and

[00236] determine a second attention feature attended by each information unit of the

first modal information to the second modal information according to the attention weight and the second modal feature.

[00237] In a possible implementation mode, the determination submodule includes a first attention determination unit and a second threshold determination unit.

[00238] The first attention determination unit is configured to determine a first attention feature attended by the second modal information to the first modal information according to the modal feature of the first modal information and the modal feature of the second modal information.

[00239] The second threshold determination unit is configured to determine a second fusion threshold parameter corresponding to the second modal information according to the modal feature of the second modal information and the first attention feature.

[00240] In a possible implementation mode, the first modal information includes the at least one information unit, and the second modal information includes the at least one information unit. The first attention determination unit is specifically configured to:

[00241] acquire the first modal feature of each information unit of the first modal information,

[00242] acquire the second modal feature of each information unit of the second modal information,

[00243] determine the attention weight between each information unit of the first modal information and each information unit of the second modal information according to the first modal feature and the second modal feature and

[00244] determine a first attention feature attended by each information unit of the second modal information to the first modal information according to the attention weight and the first modal feature.

[00245] In a possible implementation mode, the fusion submodule includes a second attention determination unit and a first fusion unit.

- [00246] The second attention determination unit is configured to determine the second attention feature attended by the first modal information to the second modal information according to the modal feature of the first modal information and the modal feature of the second modal information.
- [00247] The first fusion unit is configured to perform feature fusion on the modal feature of the first modal information and the second attention feature by using the fusion threshold parameter to determine the first fused feature corresponding to the first modal information.
- [00248] In a possible implementation mode, the first fusion unit is specifically configured to:
- [00249] perform feature fusion on the modal feature of the first modal information and the second attention feature to obtain a first fusion result;
- [00250] process the first fusion result by using the fusion threshold parameter to obtain a processed first fusion result; and
- [00251] determine the first fused feature corresponding to the first modal information based on the processed first fusion result and the first modal feature.
- [00252] In a possible implementation mode, the fusion submodule includes a first attention determination unit and a second fusion unit.
- [00253] The first attention determination unit is configured to determine the first attention feature attended by the second modal information to the first modal information according to the modal feature of the first modal information and the modal feature of the second modal information.
- [00254] The second fusion unit is configured to determine the second fused feature corresponding to the second modal information according to the modal feature of the second modal information and the first attention feature.
- [00255] In a possible implementation mode, the second fusion unit is specifically configured to:

- [00256] perform feature fusion on the modal feature of the second modal information and the first attention feature to obtain a second fusion result;
- [00257] process the second fusion result by using the fusion threshold parameter to obtain a processed second fusion result; and
- [00258] determine the second fused feature corresponding to the second modal information based on the processed second fusion result and the second modal feature.
- [00259] In a possible implementation mode, the determination module 83 is specifically configured to:
- [00260] determine the similarity between the first modal information and the second modal information based on first attention information of the first fused feature and second attention information of the second fused feature.
- [00261] In a possible implementation mode, the first modal information is information to be retrieved of a first modality, and the second modal information is pre-stored information of a second modality; and the device further includes a retrieval result determination module.
- [00262] The retrieval result determination module is configured to determine the second modal information as a retrieval result of the first modal information in condition that the similarity meets a preset condition.
- [00263] In a possible implementation mode, there are multiple pieces of second modal information, and the retrieval result determination module includes a sequencing submodule, an information determination submodule and a retrieval result determination submodule.
- [00264] The sequencing submodule is configured to sequence the multiple pieces of second modal information according to a similarity between the first modal information and each piece of second modal information to obtain a sequencing result.
- [00265] The information determination submodule is configured to determine the second modal information that the similarity meets the preset condition according to the

sequencing result.

[00266] The retrieval result determination submodule is configured to determine the second modal information that the similarity meets the preset condition as the retrieval result of the first modal information.

[00267] In a possible implementation mode, the preset condition includes any one of the following conditions.

[00268] The similarity is greater than a preset value; and a rank of the similarity sequenced from low to high is higher than a preset rank.

[00269] In a possible implementation mode, the first modal information includes one piece of modal information in text information or image information; and the second modal information includes the other piece of modal information in the text information or the image information.

[00270] In a possible implementation mode, the first modal information is training sample information of the first modality, the second modal information is training sample information of the second modality, and each piece of training sample information of the first modality and each piece of training sample information of the second modality form a training sample pair.

[00271] In a possible implementation mode, the training sample pair includes a positive sample pair and a negative sample pair. The device further includes a feedback module, configured to:

[00272] acquire a similarity of each training sample pair,

[00273] determine loss in feature fusion of the first modal information and the second modal information according to the similarity of the positive sample pair with the highest modal information matching degree in the positive sample pairs and the similarity of the negative sample pair with the lowest matching degree in the negative sample pairs and

[00274] adjust a model parameter of a cross-modal information retrieval model adopted for the feature fusion process of the first modal information and the second modal

information according to the loss.

[00275] It can be understood that various method embodiments as mentioned above in the disclosure may be combined to form combined embodiments without departing from principles and logics. For saving the space, elaborations are omitted in the disclosure.

[00276] In addition, the present disclosure also provides the abovementioned device, an electronic device, a computer-readable storage medium and a program. All of them may be configured to implement any cross-modal information retrieval method provided in the disclosure. Corresponding technical solutions and descriptions refer to the corresponding records in the method embodiments and are not be elaborated.

[00277] FIG. 9 is a block diagram of a cross-modal information retrieval device 1900 for cross-modal information retrieval according to an exemplary embodiment of the present disclosure. For example, the device 1900 may be provided as a server. Referring to Fig. 9, the device 1900 includes a processing component 1922, further including one or more processors, and memory resources represented by a memory 1932, configured to store instructions executable by the processing component 1922, for example, an application program. The application program stored in the memory 1932 may include one or more modules, each of which corresponds to a set of instructions. In addition, the processing component 1922 is configured to execute the instructions to implement the abovementioned method.

[00278] The device 1900 may further include a power component 1926 configured to perform power management of the device 1900, a wired or wireless network interface 1950 configured to connect the device 1900 to a network, and an Input/Output (I/O) interface 1958. The device 1900 may operate based on an operating system stored in the memory 1932, for example, Windows Server™, Mac OS X™, Unix™, Linux™, FreeBSD™ or the like.

[00279] In an exemplary embodiment, a non-transitory computer-readable storage medium is also provided, which includes, for example a memory 1932 including computer program instructions. The computer program instructions may be executed by the processing component 1922 of the device 1900 to implement the abovementioned

method.

[00280] The present disclosure may be a system, a method and/or a computer program product. The computer program product may include a computer-readable storage medium, which stores computer-readable program instructions configured to enable a processor to implement various aspects of the present disclosure.

[00281] The computer-readable storage medium may be a tangible device capable of retaining and storing instructions used by an instruction execution device. For example, the computer-readable storage medium may be, but not limited to, an electric storage device, a magnetic storage device, an optical storage device, an electromagnetic storage device, a semiconductor storage device or any appropriate combination thereof. More specific examples (non-exhaustive list) of the computer-readable storage medium include a portable computer disk, a hard disk, a Random Access Memory (RAM), a Read-Only Memory (ROM), an Erasable Programmable ROM (EPROM) (or a flash memory), a Static RAM (SRAM), a Compact Disc Read-Only Memory (CD-ROM), a Digital Video Disk (DVD), a memory stick, a floppy disk, a mechanical coding device, a punched card or in-slot raised structure with instructions stored therein, and any appropriate combination thereof. Herein, the computer-readable storage medium is not explained as a transient signal, for example, a radio wave or another freely propagated electromagnetic wave, an electromagnetic wave propagated through a wave guide or another transmission medium (for example, a light pulse propagated through an optical fiber cable) or an electric signal transmitted through an electric wire.

[00282] The computer-readable program instructions described in the disclosure may be downloaded from the computer-readable storage medium to each computing/processing device or downloaded to an external computer or an external storage device through a network such as the Internet, a Local Area Network (LAN), a Wide Area Network (WAN) and/or a wireless network. The network may include a copper transmission cable, optical fiber transmission, wireless transmission, a router, a firewall, a switch, a gateway computer and/or an edge server. A network adapter card or network interface in each computing/processing device receives the computer-readable program instructions from the network and forwards the computer-readable program

instructions for storage in the computer-readable storage medium in each computing/processing device.

[00283] The computer program instructions configured to execute the operations of the disclosure may be an assembly instruction, an Instruction Set Architecture (ISA) instruction, a machine instruction, a machine related instruction, a microcode, a firmware instruction, state setting data or a source code or target code edited by one or any combination of more programming languages, the programming language including an object-oriented programming language such as Smalltalk and C++ and a conventional procedural programming language such as "C" language or a similar programming language. The computer-readable program instructions may be completely executed in a computer of a user or partially executed in the computer of the user, executed as an independent software package, executed partially in the computer of the user and partially in a remote computer, or executed completely in the remote server or a server. Under the condition that the remote computer is involved, the remote computer may be connected to the computer of the user through any type of network including an LAN or a WAN, or, may be connected to an external computer (for example, connected by an Internet service provider through the Internet). In some embodiments, an electronic circuit such as a programmable logic circuit, a Field-Programmable Gate Array (FPGA) or a Programmable Logic Array (PLA) may be customized by use of state information of a computer-readable program instruction, and the electronic circuit may execute the computer-readable program instructions, thereby implementing various aspects of the disclosure.

[00284] Various aspects of the disclosure are described with reference to flowcharts and/or block diagrams of the method, device (system) and computer program product according to the embodiments of the disclosure. It is to be understood that each block in the flowcharts and/or the block diagrams and a combination of each block in the flowcharts and/or the block diagrams may be implemented by computer-readable program instructions.

[00285] These computer-readable program instructions may be provided to a universal computer, a dedicated computer or a processor of another programmable data processing

device, thereby generating a machine to further generate a device that realizes a function/action specified in one or more blocks in the flowcharts and/or the block diagrams when the instructions are executed through the computer or the processor of the other programmable data processing device. These computer-readable program instructions may also be stored in a computer-readable storage medium, and enable the computer, the programmable data processing device and/or another device to operate in a specific manner, so that the computer-readable medium including the instructions includes a product including instructions for implementing each aspect of the function/action specified in one or more blocks of the flowcharts and/or the block diagrams.

[00286] These computer-readable program instructions may further be loaded to the computer, the other programmable data processing device or the other device, so that a series of operating steps are executed in the computer, the other programmable data processing device or the other device to generate a process implemented by the computer to further realize the functions/actions specified in one or more blocks in the flowcharts and/or the block diagrams by the instructions executed in the computer, the other programmable data processing device or the other device.

[00287] The flowcharts and block diagrams in the drawings illustrate system architectures, functions and operations of the system, method and computer program product that may be realized according to multiple embodiments of the disclosure. In this regard, each block in the flowcharts or the block diagrams may represent a module, a program segment or a part of instructions, and part of the module, which includes one or more executable instructions for implementing the specified logical functions. In some alternative implementations, the functions marked in the blocks may also be realized in a sequence different from those marked in the drawings. For example, two continuous blocks can actually be executed substantially concurrently, or may also be executed in a reverse order sometimes, which depends upon the functions involved. It is further to be noted that each block in the block diagrams and/or the flowcharts and a combination of the blocks in the block diagrams and/or the flowcharts can be implemented by a dedicated hardware-based system for implementing specified functions or operations, or by a combination of dedicated hardware and computer instructions.

[00288] The forgoing has described each embodiment of the disclosure, which are exemplary but are not intended to be exhaustive, and also are not limited to each embodiment disclosed. Many modifications and variations are apparent to those of ordinary skill in the art without departing from the scope and spirit of each described embodiment of the disclosure. The terms used herein are selected to explain the principle and practical application of each embodiment or technical improvements in the technologies in the market best or enable others of ordinary skill in the art to understand each embodiment disclosed herein.

CLAIMS

1. A cross-modal information retrieval method, comprising:

acquiring first modal information and second modal information;

performing feature fusion on a modal feature of the first modal information and a modal feature of the second modal information to determine a first fused feature corresponding to the first modal information and a second fused feature corresponding to the second modal information; and

determining a similarity between the first modal information and the second modal information based on the first fused feature and the second fused feature.

2. The method of claim 1, wherein performing feature fusion on the modal feature of the first modal information and the modal feature of the second modal information to determine the first fused feature corresponding to the first modal information and the second fused feature corresponding to the second modal information comprises:

determining, based on the modal feature of the first modal information and the modal feature of the second modal information, a fusion threshold parameter for feature fusion of the first modal information and the second modal information; and

performing feature fusion on the modal feature of the first modal information and the modal feature of the second modal information based on the fusion threshold parameter to determine the first fused feature corresponding to the first modal information and the second fused feature corresponding to the second modal information, wherein the fusion threshold parameter is configured for fused features obtained by feature fusion according to a matching degree between features, and the fusion threshold parameter becomes smaller as the matching degree between the features is lower.

3. The method of claim 2, wherein determining, based on the modal feature of the first modal

information and the modal feature of the second modal information, the fusion threshold parameter for feature fusion of the first modal information and the second modal information comprises:

determining a second attention feature attended by the first modal information to the second modal information according to the modal feature of the first modal information and the modal feature of the second modal information; and

determining a first fusion threshold parameter corresponding to the first modal information according to the modal feature of the first modal information and the second attention feature.

4. The method of claim 3, wherein the first modal information comprises at least one information unit, and the second modal information comprises at least one information unit; and wherein determining the second attention feature attended by the first modal information to the second modal information comprises:

acquiring a first modal feature of each of the at least one information unit of the first modal information;

acquiring a second modal feature of each of the at least one information unit of the second modal information;

determine an attention weight between each information unit of the first modal information and each information unit of the second modal information according to the first modal feature and the second modal feature; and

determining a second attention feature attended by each information unit of the first modal information to the second modal information according to the attention weight and the second modal feature.

5. The method of claim 2, wherein determining, based on the modal feature of the first modal information and the modal feature of the second modal information, the fusion threshold

parameter for feature fusion of the first modal information and the second modal information comprises:

determining a first attention feature attended by the second modal information to the first modal information according to the modal feature of the first modal information and the modal feature of the second modal information; and

determining a second fusion threshold parameter corresponding to the second modal information according to the modal feature of the second modal information and the first attention feature.

6. The method of claim 5, wherein the first modal information comprises at least one information unit, and the second modal information comprises at least one information unit; and wherein determining the first attention feature attended by the second modal information to the first modal information according to the modal feature of the first modal information and the modal feature of the second modal information comprises:

acquiring the first modal feature of each of the at least one information unit of the first modal information;

acquiring the second modal feature of each of the at least one information unit of the second modal information;

determining an attention weight between each information unit of the first modal information and each information unit of the second modal information according to the first modal feature and the second modal feature; and

determining a first attention feature attended by each information unit of the second modal information to the first modal information according to the attention weight and the first modal feature.

7. The method of claim 2, wherein determining the first fused feature corresponding to the first modal information comprises:

determining a second attention feature attended by the first modal information to the second modal information according to the modal feature of the first modal information and the modal feature of the second modal information; and

performing feature fusion on the modal feature of the first modal information and the second attention feature by using the fusion threshold parameter to determine the first fused feature corresponding to the first modal information.

8. The method of claim 7, wherein performing feature fusion on the modal feature of the first modal information and the second attention feature by using the fusion threshold parameter to determine the first fused feature corresponding to the first modal information comprises:

performing feature fusion on the modal feature of the first modal information and the second attention feature to obtain a first fusion result;

processing, by using the fusion threshold parameter, the first fusion result to obtain a processed first fusion result; and

determining the first fused feature corresponding to the first modal information based on the processed first fusion result and the first modal feature.

9. The method of claim 2, wherein determining the second fused feature corresponding to the second modal information comprises:

determining a first attention feature attended by the second modal information to the first modal information according to the modal feature of the first modal information and the modal feature of the second modal information; and

determining the second fused feature corresponding to the second modal information according to the modal feature of the second modal information and the first attention feature.

10. The method of claim 9, wherein determining the second fused feature corresponding to the second modal information according to the modal feature of the second modal information

and the first attention feature comprises:

performing feature fusion on the modal feature of the second modal information and the first attention feature to obtain a second fusion result;

processing, by using the fusion threshold parameter, the second fusion result to obtain a processed second fusion result; and

determining the second fused feature corresponding to the second modal information based on the processed second fusion result and the second modal feature.

11. The method of claim 1, wherein determining the similarity between the first modal information and the second modal information based on the first fused feature and the second fused feature comprises:

determining the similarity between the first modal information and the second modal information based on first attention information of the first fused feature and second attention information of the second fused feature.

12. The method of claim 1, wherein the first modal information comprises information to be retrieved of a first modality, and the second modal information comprises pre-stored information of a second modality; and wherein the method further comprises:

determining the second modal information as a retrieval result of the first modal information in condition that the similarity meets a preset condition.

13. The method of claim 12, wherein the second modal information comprises multiple pieces of second modal information, and wherein determining the second modal information as the retrieval result of the first modal information in condition that the similarity meets the preset condition comprises:

sequencing the multiple pieces of second modal information according to a similarity between the first modal information and each of the multiple pieces of second modal

information to obtain a sequencing result;

determining, according to the sequencing result, a second modal information that the similarity meets the preset condition; and

determining the second modal information that the similarity meets the preset condition as the retrieval result of the first modal information.

14. The method of claim 13, wherein the preset condition comprises any one of the following conditions that:

the similarity is greater than a preset value; and a rank of the similarity sequenced from low to high is higher than a preset rank.

15. The method of claim 1, wherein the first modal information comprises one piece of modal information in text information or image information; and the second modal information comprises the other piece of modal information in the text information or the image information.

16. The method of claim 1, wherein the first modal information comprises training sample information of a first modality, the second modal information comprises training sample information of a second modality, and wherein each piece of the training sample information of the first modality and each piece of the training sample information of the second modality form a training sample pair.

17. The method of claim 16, wherein the training sample pair comprises a positive sample pair and a negative sample pair; and wherein the method further comprises:

acquiring a similarity of each training sample pair,

determining a loss in feature fusion of the first modal information and the second modal information according to a similarity of a positive sample pair with a highest matching

degree of modal information in positive sample pairs and a similarity of a negative sample pair with a lowest matching degree in negative sample pairs, and

adjusting, according to the loss, a model parameter of a cross-modal information retrieval model that is adopted for the feature fusion of the first modal information and the second modal information.

18. A cross-modal information retrieval device, comprising:

an acquisition module, configured to acquire first modal information and second modal information;

a fusion module, configured to perform feature fusion on a modal feature of the first modal information and a modal feature of the second modal information to determine a first fused feature corresponding to the first modal information and a second fused feature corresponding to the second modal information; and

a determination module, configured to determine a similarity between the first modal information and the second modal information based on the first fused feature and the second fused feature.

19. The device of claim 18, wherein the fusion module comprises:

a determination submodule, configured to determine, based on the modal feature of the first modal information and the modal feature of the second modal information, a fusion threshold parameter for feature fusion of the first modal information and the second modal information; and

a fusion submodule, configured to perform feature fusion on the modal feature of the first modal information and the modal feature of the second modal information based on the fusion threshold parameter to determine the first fused feature corresponding to the first modal information and the second fused feature corresponding to the second modal information, wherein the fusion threshold parameter is configured for fused features obtained

by feature fusion according to a matching degree between features, and the fusion threshold parameter becomes smaller as the matching degree between the features is lower.

20. The device of claim 19, wherein the determination submodule comprises:

a second attention determination unit, configured to determine a second attention feature attended by the first modal information to the second modal information according to the modal feature of the first modal information and the modal feature of the second modal information; and

a first threshold determination unit, configured to determine a first fusion threshold parameter corresponding to the first modal information according to the modal feature of the first modal information and the second attention feature.

21. The device of claim 20, wherein the first modal information comprises at least one information unit, and the second modal information comprises at least one information unit; and wherein the second attention determination unit is further configured to:

acquire a first modal feature of each of the at least one information unit of the first modal information;

acquire a second modal feature of each of the at least one information unit of the second modal information;

determine an attention weight between each information unit of the first modal information and each information unit of the second modal information according to the first modal feature and the second modal feature; and

determine a second attention feature attended by each information unit of the first modal information to the second modal information according to the attention weight and the second modal feature.

22. The device of claim 19, wherein the determination submodule comprises:

a first attention determination unit, configured to determine a first attention feature attended by the second modal information to the first modal information according to the modal feature of the first modal information and the modal feature of the second modal information; and

a second threshold determination unit, configured to determine a second fusion threshold parameter corresponding to the second modal information according to the modal feature of the second modal information and the first attention feature.

23. The device of claim 22, wherein the first modal information comprises at least one information unit, and the second modal information comprises at least one information unit; and wherein the first attention determination unit is further configured to:

acquire the first modal feature of each of the at least one information unit of the first modal information;

acquire the second modal feature of each of the at least one information unit of the second modal information;

determine an attention weight between each information unit of the first modal information and each information unit of the second modal information according to the first modal feature and the second modal feature; and

determine a first attention feature attended by each information unit of the second modal information to the first modal information according to the attention weight and the first modal feature.

24. The device of claim 19, wherein the fusion submodule comprises:

a second attention determination unit, configured to determine a second attention feature attended by the first modal information to the second modal information according to the modal feature of the first modal information and the modal feature of the second modal information; and

a first fusion unit, configured to perform feature fusion on the modal feature of the first modal information and the second attention feature by using the fusion threshold parameter to determine the first fused feature corresponding to the first modal information.

25. The device of claim 24, wherein the first fusion unit is further configured to:

perform feature fusion on the modal feature of the first modal information and the second attention feature to obtain a first fusion result;

process, by using the fusion threshold parameter, the first fusion result to obtain a processed first fusion result; and

determine the first fused feature corresponding to the first modal information based on the processed first fusion result and the first modal feature.

26. The device of claim 19, wherein the fusion submodule comprises:

a first attention determination unit, configured to determine a first attention feature attended by the second modal information to the first modal information according to the modal feature of the first modal information and the modal feature of the second modal information; and

a second fusion unit, configured to determine the second fused feature corresponding to the second modal information according to the modal feature of the second modal information and the first attention feature.

27. The device of claim 26, wherein the second fusion unit is further configured to:

perform feature fusion on the modal feature of the second modal information and the first attention feature to obtain a second fusion result;

process, by using the fusion threshold parameter, the second fusion result to obtain a processed second fusion result; and

determine the second fused feature corresponding to the second modal information based on the processed second fusion result and the second modal feature.

28. The device of claim 18, wherein the determination module is further configured to:

determine the similarity between the first modal information and the second modal information based on first attention information of the first fused feature and second attention information of the second fused feature.

29. The device of claim 18, wherein the first modal information comprises information to be retrieved of a first modality, and the second modal information is pre-stored information of a second modality; and wherein the device further comprises:

a retrieval result determination module, configured to determine the second modal information as a retrieval result of the first modal information in condition that the similarity meets a preset condition.

30. The device of claim 29, wherein the second modal information comprises multiple pieces of second modal information, and wherein the retrieval result determination module comprises:

a sequencing submodule, configured to sequence the multiple pieces of second modal information according to a similarity between the first modal information and each of the multiple pieces of second modal information to obtain a sequencing result;

an information determination submodule, configured to determine, according to the sequencing result, the second modal information that the similarity meets the preset condition; and

a retrieval result determination submodule, configured to determine the second modal information that the similarity meets the preset condition as the retrieval result of the first modal information.

31. The device of claim 30, wherein the preset condition comprises any one of the following conditions that:

the similarity is greater than a preset value; and a rank of the similarity from low to high is higher than a preset rank.

32. The device of claim 18, wherein the first modal information comprises one piece of modal information in text information or image information; and the second modal information comprises the other piece of modal information in the text information or the image information.

33. The device of claim 18, wherein the first modal information comprises training sample information of a first modality, the second modal information comprises training sample information of a second modality, and wherein each piece of the training sample information of the first modality and each piece of the training sample information of the second modality form a training sample pair.

34. The device of claim 33, wherein the training sample pairs comprise positive sample pairs and negative sample pairs; and wherein the device further comprises a feedback module, configured to:

acquire a similarity of each training sample pair,

determine loss in feature fusion of the first modal information and the second modal information according to a similarity of a positive sample pair with a highest matching degree of modal information in positive sample pairs and a similarity of a negative sample pair with a lowest matching degree in negative sample pairs, and

adjust, according to the loss, a model parameter of a cross-modal information retrieval model that is adopted for the feature fusion process of the first modal information and the second modal information.

35. A cross-modal information retrieval device, comprising:

a processor; and

a memory, configured to store instructions executable by the processor,

wherein the processor is configured to execute the executable instructions stored in the memory to implement the method of any one of claims 1-17.

36. A non-transitory computer-readable storage medium, having stored therein computer program instructions that, when being executed by a processor, cause the processor to implement the method of any one of claims 1-17.