

DATA PROCESSING METHOD AND APPARATUS BASED ON MULTI-MODAL MODEL, AND DEVICE AND MEDIUM

TECHNICAL FIELD

[0001]The present application relates to the field of artificial intelligence technology, and in particular, to a data processing method, an apparatus, a device and a medium based on a multi-modal model.

BACKGROUND

[0002]With the rapid development of network technology, there are more and more network data, and in order to retrieve data required by a user from an enormous volume of network data, the user needs to provide a keyword, and a server retrieves the data required by the user from the enormous volume of network data based on the keyword. However, because the volume of network data is enormous, the server needs to consume numerous computing resources and spend much time to retrieve the data required by the user, in which way, the numerous resources of the server are consumed, and the user waits for a long time to obtain the required data. For example, many cameras may be deployed in an actual scenario. These cameras may collect a great number of images. In order to analyze the data required by the user from these images, the server needs to consume numerous computing resources and spend much time.

SUMMARY

[0003]The present application provides a data processing method based on a multi-modal model, including: determining an initial question feature matching a configured question template based on an original input question; acquiring a plurality of additional question features associated with the initial question feature based on the initial question feature; generating a plurality of target questions based on the initial question feature and the plurality of additional question features, each of the plurality of target questions including the initial question feature and at least one of the plurality of additional question features; for each of the plurality of target questions, inputting the target question into a trained multi-modal target model, and outputting, by the multi-modal target model, candidate answers corresponding to the target question; determining a target answer corresponding to the original input question based on the candidate answers corresponding to each of the plurality of target questions.

[0004]The present application provides a data processing apparatus based on a multi-modal model, including: a determining module, configured to determine an initial question feature

matching a question template based on an original input question; an acquiring module, configured to acquire a plurality of additional question features associated with the initial question feature; a generating module, configured to generate a plurality of target questions based on the initial question feature and the plurality of additional question features, each of the plurality of target questions including the initial question feature and at least one of the plurality of additional question features, where the acquiring module is further configured to, for each of the plurality of target questions, input the target question into a trained multi-modal target model, and output, by the multi-modal target model, candidate answers corresponding to the target question; the determining module is further configured to determine a target answer corresponding to the original input question based on the candidate answers corresponding to each of the plurality of target questions, the target answer being an output result of the original input question.

[0005]The present application provides an electronic device, including: a processor, and a machine readable storage medium storing machine executable instructions that can be executed by the processor, where the processor is configured to execute the machine executable instructions to implement the data processing method based on the multi-modal model as exemplified.

[0006]As can be seen from the above technical solutions, in the embodiments of the present application, the plurality of target questions can be generated based on the initial question feature and the plurality of additional question features, the candidate answers corresponding to each target question is obtained by using the multi-modal target model, and the target answer corresponding to the original input question is determined based on the candidate answers corresponding to each target question, such that the target answer is obtained with the help of the multi-modal target model. Moreover, an accurate and reliable target answer is obtained through the plurality of target questions, and the target answer corresponding to the original input question can be searched for from an enormous volume of data only by consuming a small amount of computing resources, in which way, time for acquiring the target answer is shorter, saving computing resources of a server and time consumption, quickly locating the target answer corresponding to the original input question, providing abnormality logic analysis and reasoning capabilities, implementing automatic abnormality diagnosis, giving the accurate and reliable target answer through an analysis capability of the multi-modal target model, and comprehensively considering possibilities of various dimensions such as an image, a video and a voice.

BRIEF DESCRIPTION OF THE DRAWINGS

[0007]In order to more clearly illustrate the technical solutions in the embodiments of the present application or in the related art, the accompanying drawings required for description of the embodiments of the present application or the related art will be briefly introduced below. The drawings in the following description are merely some embodiments recited in the application. For those skilled in the art, other drawings may be acquired according to these drawings in the embodiments of this application.

[0008]FIG. 1 is a schematic flowchart of a data processing method based on a multi-modal model in an embodiment.

[0009]FIG. 2 is a schematic flowchart of a data processing method based on a multi-modal model in an embodiment.

[0010]FIG. 3 is a schematic diagram of a self-reasoning method based on a multi-modal model in an embodiment.

[0011]FIG. 4 is a schematic structure diagram of a data processing apparatus based on a multi-modal model in an embodiment.

[0012]FIG. 5 is a hardware structure diagram of an electronic device in an embodiment.

DETAILED DESCRIPTION OF THE EMBODIMENTS

[0013]The terms used in the embodiments of the present application are for the purpose of describing particular examples only, and are not intended to limit the present application. Terms determined by “a,” “the” and “said” in their singular forms in the specification and claims of the present application are also intended to include plurality, unless clearly indicated otherwise in the context. It should also be understood that the term “and/or” as used herein refers to and includes any or all possible combinations of one or more of the associated listed items.

[0014] It is to be understood that, although terms “first,” “second,” “third,” and the like may be used in the embodiments of the present application to describe various information, such information should not be limited to these terms. These terms are only used to distinguish one category of information from another. For example, without departing from the scope of the present application, first information may be referred as second information; and similarly, second information may also be referred as first information. Depending on the context, the word “if” as used herein may be interpreted as “when” or “upon” or “in response to determining”.

[0015]An embodiment of the present application provides a data processing method based on a multi-modal model. The method may be applied to any device, for example, a server. As

shown in FIG. 1, the method may include:

[0016]At step 101, an initial question feature matching a configured question template is determined based on an original input question.

[0017]At step 102, a plurality of additional question features associated with the initial question feature are acquired based on the initial question feature.

[0018]At step 103, a plurality of target questions are generated based on the initial question feature and the plurality of additional question features, where each target question includes the initial question feature and at least one additional question feature.

[0019]At step 104, for each target question, the target question is input into a trained multi-modal target model, and candidate answers corresponding to the target question are output by the multi-modal target model.

[0020]At step 105, a target answer corresponding to the original input question is determined based on the candidate answers corresponding to each target question, and the target answer may be a final output result of the original input question

[0021]In some embodiments, determining the initial question feature matching the configured question template based on the original input question may include but is not limited to: if the question template includes a feature type, determining whether a key value matching the feature type exists in the original input question; if the key value matching the feature type exists, determining the initial question feature based on the key value; if the key value matching the feature type does not exist, analyzing the original input question to obtain an initial question feature matching the feature type; or, sending prompt information for the feature type to a user, a new original input question is provided by the user based on the prompt information, and if the key value matching the feature type exists in the new original input question, determining the initial question feature based on the key value.

[0022]In some embodiments, acquiring the plurality of additional question features associated with the initial question feature based on the initial question feature may include but is not limited to: inputting the initial question feature into a trained feature analysis target model, and outputting the plurality of additional question features associated with the initial question feature by the feature analysis target model. Or, the initial question feature is input into the multi-modal target model, and the plurality of additional question features associated with the initial question feature are output by the multi-modal target model. Or, the plurality of additional question features associated with the initial question feature are analyzed based on the initial question feature (that is, the initial question feature is analyzed to obtain the plurality of additional question features associated with the initial question feature).

[0023]In some embodiments, the feature analysis target model may be a feature analysis target model matching a target scenario type; the question template may be a question template matching the target scenario type, where the target scenario type may be a behavior abnormality type, a vehicle abnormality type, or an industry abnormality type.

[0024]In some embodiments, generating the plurality of target questions based on the initial question feature and the plurality of additional question features may include but is not limited to: selecting a candidate additional question feature from the plurality of additional question features; after the initial question feature and the candidate additional question feature are combined, for each combination, adding context information to the initial question feature and/or the candidate additional question feature to obtain a target question corresponding to the candidate additional question feature.

[0025]In some embodiments, determining the target answer corresponding to the original input question based on the candidate answers corresponding to each target question may include but is not limited to: filtering candidate answers with repeated contents from all candidate answers based on the candidate answers corresponding to each target question to obtain at least one remaining candidate answer, and determining a number of occurrences of each remaining candidate answer in all the candidate answers; determining a remaining candidate answer with a maximum number of occurrences as the target answer corresponding to the original input question; or, if a ratio value of occurrences of a remaining candidate answer corresponding to the maximum number of occurrences in all the candidate answers is greater than a preset threshold, determining the remaining candidate answer with the maximum number of occurrences as the target answer corresponding to the original input question.

[0026]In some embodiments, after determining the target answer corresponding to the original input question based on the candidate answers corresponding to each target question, target data corresponding to the target answer may be output, and the target data may include the target answer, or the target data may include the target answer and additional question features corresponding to the target answer, where the target answer may include but is not limited to an answer of at least one dimension of an image, a text, a sound, or a video; where the target answer may correspond to a plurality of candidate answers, and the additional question features corresponding to the target answer may include an additional question feature corresponding to each candidate answer corresponding to the target answer.

[0027]In some embodiments, the above execution sequence is merely for the convenience of describing the given examples. In actual applications, the execution sequence between the steps may be changed. Therefore, the execution sequence is not limited. Moreover, in other

embodiments, steps in corresponding methods are not necessarily executed in the order shown and described in the present application, and the steps included in the methods in other embodiments may be more or less than those described in this application. In addition, a single step described in the application may be decomposed into multiple steps for description in other embodiments; multiple steps described in the application may be combined into a single step for description in other embodiments.

[0028]As can be seen from the above technical solutions, in the embodiments of the present application, the plurality of target questions can be generated based on the initial question feature and the plurality of additional question features, the candidate answers corresponding to each target question is obtained by using the multi-modal target model, and the target answer corresponding to the original input question is determined based on the candidate answers corresponding to each target question, such that the target answer is obtained with the help of the multi-modal target model. Moreover, an accurate and reliable target answer is obtained through the plurality of target questions, and the target answer corresponding to the original input question can be searched for from an enormous volume of data only by consuming a small amount of computing resources, in which way, time for acquiring the target answer is shorter, saving computing resources of a server and time consumption, quickly locating the target answer corresponding to the original input question, providing abnormality logic analysis and reasoning capabilities, implementing automatic abnormality diagnosis, giving the accurate and reliable target answer through an analysis capability of the multi-modal target model, and comprehensively considering possibilities of various dimensions such as an image, a video and a voice.

[0029]The above technical solutions in the embodiments of the present application will be described below with reference to specific application scenarios.

[0030]Because the volume of network data (such as text data, video data, audio data, and image data) is large, the server needs to consume numerous computing resources and spend much time to retrieve the data required by the user, in which way, the numerous resources of the server are consumed, and the user waits for a long time to obtain the required data. For example, many cameras may be deployed in an actual scenario. These cameras may collect a great number of images. In order to analyze the data required by the user from these images, the server needs to consume numerous computing resources and spend much time.

[0031]For the above findings, an embodiment of the present application provides a data processing method based on a multi-modal model. According to the method, the target answer is obtained with the help of the multi-modal target model. Moreover, an accurate and reliable

target answer is obtained through the plurality of target questions, and the target answer corresponding to the original input question can be searched for from an enormous volume of data only by consuming a small amount of computing resources, in which way, time for acquiring the target answer is shorter, saving computing resources of a server and time consumption, and quickly locating the target answer corresponding to the original input question

[0032]In some embodiments, a question template may be pre-configured, and the question template may include at least one feature type, indicating an original input question should provide a key value matching the feature type. The question template may be a question template matching a scenario type. For example, the scenario type may include but is not limited to a behavior abnormality type, a vehicle abnormality type, an industry abnormality type, or the like. Therefore, a question template a1 matching the behavior abnormality type, a question template a2 matching the vehicle abnormality type, a question template a3 matching the industry abnormality type, or the like may be pre-configured. Of course, the above types are only examples of the scenario type, which is not limited.

[0033]For example, the question template a1 may include but is not limited to at least one of the following: a time type, a position type, or a behavior event type. The behavior event type may be a specific behavior, for example, a fighting behavior, a robbing behavior, or a lost item behavior, and the behavior event type is subsequently denoted as a behavior X, indicating an event type in which the behavior X occurs, which is not limited.

[0034]The question template a2 may include but is not limited to at least one of the following: a time type, a position type, or a behavior event type. The behavior event type may be a specific vehicle behavior, for example, a “red light violation” behavior, or an “overspeed” behavior, indicating an event type in which the specific vehicle behavior occurs, which is not limited.

[0035]The question template a3 may include but is not limited to at least one of the following: a sound type, a vibration type, an acceleration type, an angular velocity type, a temperature type, or a humidity type, which is not limited.

[0036]Of course, the above templates are only some examples of the question template, which is not limited.

[0037]In some embodiments, for each scenario type, a feature analysis target model corresponding to the scenario type may be trained. For example, a feature analysis target model b1 corresponding to the behavior abnormality type, a feature analysis target model b2 corresponding to the vehicle abnormality type, and a feature analysis target model b3 corresponding to the industry abnormality type may be trained. Of course, a same feature

analysis target model may be trained for all scenario types.

[0038]In order to train the feature analysis target model b1, sample data related to the “behavior abnormality type” may be collected. The sample data includes information related to the time type, information related to the position type, and information related to the behavior X. In addition to the above information, the sample data may include other types of information. The training may be performed based on an enormous volume of the sample data to obtain the feature analysis target model b1. The training process is not limited.

[0039]In order to train the feature analysis target model b2, sample data related to the “vehicle abnormality type” may be collected. The sample data includes information related to the time type, information related to the position type, and information related to the specific vehicle behavior. In addition to the above information, the sample data may include other types of information. The training may be performed based on an enormous volume of the sample data to obtain the feature analysis target model b2. The training process is not limited.

[0040]In order to train the feature analysis target model b3, sample data related to the “industry abnormality type” may be collected. The sample data includes information related to the sound type, information related to the vibration type, information related to the acceleration type, information related to the angular velocity type, information related to the temperature type, and information related to the humidity type. In addition to the above information, the sample data may include other types of information. The training may be performed based on an enormous volume of the sample data to obtain the feature analysis target model b3. The training process is not limited.

[0041]Of course, the above models are only some examples of the feature analysis target model, which is not limited.

[0042]In the above application scenarios, an embodiment of the present application provides a data processing method based on a multi-modal model. The method may be applied to a server. As shown in FIG. 2, the method may include following steps.

[0043]At step 201, an original input question is acquired. For example, the server may receive the original input question input by a user, or acquire the original input question in other manners. Sources of the original input question are not limited.

[0044]For example, for an application scenario of the behavior abnormality type, an example of the original input problem may be: “help me search for pedestrians appearing at an intersection of Chuangxin Road at about 3:00 pm on March 21, 2022, and of these pedestrians, who is most likely to be a target of the behavior X occurring at the intersection at this time point”.

[0045]For example, for an application scenario of the vehicle abnormality type, an example of the original input problem may be: “help me search for vehicles appearing at an intersection of Chuangxin Road at about 3:00 pm on March 21, 2022, and of these vehicles, which is most likely to be a target of the red light violation behavior occurring at the intersection at this time point”.

[0046]Of course, the above questions are only two examples of the original input problem, which is not limited.

[0047]At step 202, an initial question feature matching a question template is determined based on the original input question.

[0048]In some embodiments, the server may determine a target scenario type (that is, a scenario type of a current application scenario) based on or not based on the original input question. For example, the server may determine the target scenario type based on the original input question, determine a question template matching the target scenario type, and determine an initial question feature matching the question template based on the original input question. For example, if it is determined based on the original input question that the target scenario type is the behavior abnormality type, the server determines an initial question feature matching the question template a1 based on the original input question. The server may determine the target scenario type not based on the original input question. For example, the server may first determine the target scenario type (which may be pre-determined by a user), and determine a question template matching the target scenario type. The original input question input now needs to conform to the scenario type of the current application scenario. Then, the server determines an initial question feature matching the question template. For example, if the target scenario type is the vehicle abnormality type, the server determines an initial question feature matching the question template a2 based on the original input question. If the target scenario type is the industry abnormality type, the server determines an initial question feature matching the question template a3 based on the original input question.

[0049]For example, the question template matching the target scenario type may be the question template a1, and the question template a1 may include the time type, the position type, and the behavior event type. The server extracts, from the original input question, a key value “3:00 pm on March 21, 2022” matching the time type, a key value “intersection of Chuangxin Road” matching the position type, and a key value “behavior X” matching the behavior event type. The server determines according to these key values that the question template matching the target scenario type is the question template a1, and uses the key values as initial question features, that is, the initial question features include “3:00 pm on March 21, 2022”,

“intersection of Chuangxin Road”, and “behavior X”.

[0050]For another example, the question template matching the target scenario type may be the question template a2, and the question template a2 may include the time type, the position type, and the behavior event type. The server extracts, from the original input question, a key value “3:00 pm on March 21, 2022” matching the time type, a key value “intersection of Chuangxin Road” matching the position type, and a key value “red light violation behavior” matching the behavior event type. The server determines according to these key values that the question template matching the target scenario type is the question template a2, and may use the key values as initial question features.

[0051]In conclusion, it can be seen that the initial question feature may be extracted from the original input question based on the question template matching the target scenario type. Extraction manners of the initial question feature are not limited.

[0052]In a possible embodiment, in order to determine the initial question feature, the following manners may be used:

[0053]Manner 1: For each feature type (for example, the time type, the position type, or the behavior event type) in the question template, it is determined whether a key value matching the feature type exists in the original input question; if the key value matching the feature type exists, the initial question feature is determined based on the key value, for example, the key value is used as the initial question feature. If the key value matching the feature type does not exist, the original input question is analyzed to obtain an initial question feature matching the feature type.

[0054]For example, if the key value matching the feature type exists in the original input question, the key value may be replaced with a variable through key-value mapping, such that the key value is used as the initial question feature.

[0055]For example, if the key value matching the feature type does not exist in the original input question, the original input question may be analyzed to obtain a key value matching the feature type, and the key value is used as the initial question feature. The analyzing process of the original input question is not limited in this embodiment.

[0056]Obviously, when the question template corresponds to multiple feature types, a key value matching each feature type needs to be determined, that is, each key value is used as an initial question feature, and each feature type corresponds to one initial question feature.

[0057]Manner 2: For each feature type in the question template, it is determined whether a key value matching the feature type exists in the original input question; if the key value matching the feature type exists, the initial question feature is determined based on the key value, for

example, the key value is used as the initial question feature. If the key value matching the feature type does not exist, prompt information for the feature type is sent to a user (the prompt information indicates that the key value matching the feature type lacks in the original input question), and the user provides a new original input question based on the prompt information (that is, the key value matching the feature type is added to the previous original input question to obtain the new original input question). If the key value matching the feature type exists in the new original input question, the initial question feature may be determined based on the key value.

[0058]Manner 3: For each feature type in the question template, it is determined whether a key value matching the feature type exists in the original input question; if the key value matching the feature type exists, the initial question feature is determined based on the key value, for example, the key value is used as the initial question feature. If the key value matching the feature type does not exist, the original input question is analyzed, and if a key value matching the feature type is obtained by analyzing, the key value is used as an initial question feature matching the feature type. If the key value matching the feature type is not obtained by analyzing, prompt information for the feature type is sent to a user, and the user provides a new original input question based on the prompt information. If the key value matching the feature type exists in the new original input question, the initial question feature is determined based on the key value.

[0059]Of course, the above three manners are only examples. Acquiring manners of the initial question feature are not limited.

[0060]In conclusion, in step 202, the initial question feature matching the question template may be determined based on the original input question. Initial question features are subsequently denoted as a feature A, a feature B, and a feature C.

[0061]At step 203, a plurality of additional question features associated with the initial question feature are acquired based on the initial question feature.

[0062]In some embodiments, after the initial question feature is obtained, the plurality of additional question features may be acquired based on the initial question feature, where the additional question features are features additionally supplemented based on the initial question feature, and the additional question features may be varied and are related to the initial question feature. For example, for the application scenario of the behavior abnormality type, the initial question feature includes the “behavior X”, and the additional question features are features related to the “behavior X”, such as a pedestrian speed feature, a pedestrian wearing feature, and a pedestrian expression feature. For the application scenario of the vehicle abnormality

type, the initial question feature includes the “red light violation behavior”, and the additional question features are features related to the “red light violation behavior”, such as a vehicle speed feature, a license plate identification feature, and a vehicle color feature.

[0063]In a possible embodiment, the plurality of additional question features may be acquired in the following manners.

[0064]Manner 1: The initial question feature is input into a trained feature analysis target model, and the plurality of additional question features associated with the initial question feature are output by the feature analysis target model.

[0065]In some embodiments, the feature analysis target model is a feature analysis target model matching the target scenario type, and the target scenario type may be the behavior abnormality type, the vehicle abnormality type, or the industry abnormality type. For example, if the target scenario type is the behavior abnormality type, the initial question feature is input into the feature analysis target model b1, and the feature analysis target model b1 processes the initial question feature to obtain the plurality of additional question features associated with the initial question feature, and outputs the plurality of additional question features. If the target scenario type is the vehicle abnormality type, the initial question feature is input into the feature analysis target model b2, and the feature analysis target model b2 processes the initial question feature to obtain the plurality of additional question features associated with the initial question feature, and outputs the plurality of additional question features. If the target scenario type is the industry abnormality type, the initial question feature is input into the feature analysis target model b3, and the feature analysis target model b3 processes the initial question feature to obtain the plurality of additional question features associated with the initial question feature, and outputs the plurality of additional question features.

[0066]In some embodiments, when the training is performed based on the sample data to obtain the feature analysis target model, the sample data may include information related to the initial question feature (such as the information related to the time type, the information related to the position type, and the information related to the behavior X), and the sample data may further include other types of information (which may be used as label information of the initial question feature, that is, information reflecting the initial question feature). Therefore, through continuous learning and optimization of the feature analysis target model, the feature analysis target model can reflect a relation between the initial question feature and the additional question features. After the initial question feature is input into the trained feature analysis target model, the plurality of additional question features associated with the initial question feature can be obtained.

[0067]Manner 2: The initial question feature is input into a multi-modal target model (that is, a multi-modal target large model), and the plurality of additional question features associated with the initial question feature are output by the multi-modal target model.

[0068]In some embodiments, the multi-modal target model is obtained through continuous learning and optimization based on big data, has a powerful processing capability, and may be applicable to data processing of all scenario types, such as the behavior abnormality type, the vehicle abnormality type, and the industry abnormality type. Sources of the multi-modal target model are not limited. After the initial question feature is input into the multi-modal target model, the multi-modal target model processes the initial question feature to obtain the plurality of additional question features associated with the initial question feature, and outputs the plurality of additional question features.

[0069]Manner 3: The plurality of additional question features associated with the initial question feature are analyzed based on the initial question feature.

[0070]For example, the server may acquire original data corresponding to the initial question feature, and analyze based on the original data to obtain the plurality of additional question features associated with the initial question feature. For example, initial question features include “3:00 pm on March 21, 2022”, “intersection of Chuangxin Road”, and “behavior X”. The server may acquire one or more images generated at the intersection of Chuangxin Road at 3:00 pm on March 21, 2022, and analyze additional question features associated with the behavior X based on the images, such as the pedestrian speed feature, the pedestrian wearing feature, and the pedestrian expression feature. Of course, the above manner is only an example of obtaining the plurality of additional question features through analysis based on the original data, and is not limited, as long as the plurality of additional question features associated with the initial question feature can be obtained through analysis.

[0071]In conclusion, the plurality of additional question features associated with the initial question feature may be obtained. Of course, the manner 1, the manner 2, and the manner 3 are only examples. Acquiring manners of the additional question features are not limited.

[0072]At step 204, a plurality of target questions are generated based on the initial question feature and the plurality of additional question features, where each target question may include the initial question feature and at least one additional question feature.

[0073]For example, the server may select at least one additional question feature from the plurality of additional question features as a candidate additional question feature, and generate a target question based on the initial question feature and the candidate additional question feature. If the plurality of additional question features include an additional question feature c1,

an additional question feature c2, and an additional question feature c3, the server may generate a target question d1 based on the initial question feature and the additional question feature c1 (that is, the additional question feature c1 is used as the candidate additional question feature), generate a target question d2 based on the initial question feature and the additional question feature c2, generate a target question d3 based on the initial question feature and the additional question feature c3, generate a target question d4 based on the initial question feature, the additional question feature c1, and the additional question feature c2, generate a target question d5 based on the initial question feature, the additional question feature c1, and the additional question feature c3, and so on.

[0074]In some embodiments, generating the target question based on the initial question feature and the candidate additional question feature may include but is not limited to: combining the initial question feature and the candidate additional question feature, and after the initial question feature and the candidate additional question feature are combined, for each combination, adding context information to the initial question feature and/or the candidate additional question feature to obtain a target question corresponding to the candidate additional question feature. For example, the server may add the context information to the initial question feature, that is, before and/or after the initial question feature. Contents of the context information are not limited. For another example, the server may add the context information to the candidate additional question feature. For another example, the server may add the context information to the initial question feature and the candidate additional question feature.

[0075]Adding the context information to a question feature means taking an environment, conditions, previous information, or context of situation related to a question into consideration to enhance understanding and solving capabilities of a model for the question. Specifically, the context information here may include: pre-text information, post-text information, a surrounding environment, and previous knowledge.

[0076]The pre-text information refers to texts or data related to a current question mentioned previously. In natural language processing, this may be information in a previous sentence, a previous paragraph, or a previous document.

[0077]The post-text information refers to subsequent texts or data related to a question. This can help a model predict future events or better understand context of situation of the current question.

[0078]The surrounding environment refers to background or environment information where a question occurs. In a visual task, this may be other object, scenario, or background information related to an image. In natural language processing, this may be external

information related to a current text, such as a time, a place, or social background.

[0079]The previous knowledge refers to previous knowledge that a model has learned or been trained. This may include a corpus, a dataset, or experience that the model contacts during the training. In conclusion, adding the context information means embedding a question into broader context of situation, such that the model can understand and solve the question more comprehensively.

[0080]At step 205, for each target question, the target question is input into a trained multi-modal target model (a multi-modal target large model), and candidate answers corresponding to the target question are output by the multi-modal target model.

[0081]In some embodiments, the multi-modal target model is obtained through continuous learning and optimization based on big data, has a powerful processing capability, and may be applicable to data processing of all scenario types. Sources of the multi-modal target model are not limited. In addition to text data, multi-modal refers to including other modal data, such as an image, a video, and a voice. For example, a file has a sound and a video, and the multi-modal target model can be used to analyze whether there is a person shouting for help in the file, whether the person shouting for help is male or female, and what clothes the person shouting for help wears. Examples of the multi-modal target model include: GPT-4V (generative pre-trained transformer-4 vision), etc. After the target question is input into the multi-modal target model, the multi-modal target model processes the target question to obtain the candidate answers corresponding to the target question, and outputs the candidate answers.

[0082]In conclusion, the candidate answers corresponding to each target question, such as a candidate answer e1 corresponding to the target question d1, a candidate answer e2 corresponding to the target question d2, a candidate answer e3 corresponding to the target question d3, and a candidate answer e4 corresponding to the target question d4, may be obtained based on the multi-modal target model.

[0083]In some embodiments, for each candidate answer, the candidate answer may be a candidate answer in an image dimension, a candidate answer in a text dimension, a candidate answer in a sound dimension, a candidate answer in a video dimension, or a candidate answer in various dimensions, for example, a candidate answer in a text and video combination dimension.

[0084]At step 206, a target answer corresponding to the original input question is determined based on the candidate answers corresponding to each target question, where the target answer may be a final output result of the original input question.

[0085]In some embodiments, based on the candidate answers corresponding to each target

question, the server may filter candidate answers with repeated contents from all candidate answers to obtain at least one remaining candidate answer, and determine a number of occurrences of each remaining candidate answer in all the candidate answers. Then, based on the number of occurrences corresponding to each remaining candidate answer, the server selects the target answer from all remaining candidate answers. For example, the server determines a remaining candidate answer with a maximum number of occurrences as the target answer corresponding to the original input question; or, if a ratio value of occurrences of a remaining candidate answer corresponding to a maximum number of occurrences in all the candidate answers is greater than a preset threshold, the server determines the remaining candidate answer with the maximum number of occurrences as the target answer corresponding to the original input question.

[0086]For example, assuming that there are 6 candidate answers in total, the candidate answer e1 is “mmm”, the candidate answer e2 is “mmm”, the candidate answer e3 is “mmm”, the candidate answer e4 is “mmm”, a candidate answer e5 is “nnn”, and a candidate answer e6 is “nnn”, after the candidate answers with the repeated contents (that is, the candidate answer e2, the candidate answer e3, the candidate answer e4, and the candidate answer e6) are filtered, the obtained remaining candidate answers are “mmm” and “nnn”. Obviously, the number of occurrences of the remaining candidate answer “mmm” in all the candidate answers is 4, and the number of occurrences of the remaining candidate answer “nnn” in all the candidate answers is 2.

[0087]The server may use the remaining candidate answer “mmm” with the maximum number of occurrences as the target answer corresponding to the original input question, that is, the target answer is “mmm”. Or, the server may determine whether the ratio value of occurrences of the remaining candidate answer corresponding to the maximum number of occurrences in all the candidate answers (that is, $4/6$) is greater than the preset threshold. If the ratio value of occurrences of the remaining candidate answer corresponding to the maximum number of occurrences is greater than the preset threshold, the remaining candidate answer “mmm” with the maximum number of occurrences is used as the target answer corresponding to the original input question. If the ratio value of occurrences of the remaining candidate answer corresponding to the maximum number of occurrences is less than or equal to the preset threshold, the target answer corresponding to the original input question is not obtained.

[0088]In conclusion, it can be seen that, for the candidate answer with the maximum number of occurrences, the candidate answer is an answer the most frequently occurring, which is a consensus found from all the candidate answers and has a higher confidence level, and it can

be considered that the candidate answer is relatively correct and reliable, improving final output accuracy. Therefore, the candidate answer is used as the target answer corresponding to the original input question, and the target answer may be a final output result of the original input question.

[0089]At step 207, target data corresponding to the target answer is output, where the target data may include the target answer, or the target data may include the target answer and additional question features corresponding to the target answer.

[0090]In some embodiments, the target answer may include but is not limited to an answer of at least one dimension of an image, a text, a sound, or a video. For example, the target answer may be a target answer in an image dimension, a target answer in a text dimension, a target answer in a sound dimension, a target answer in a video dimension, or a target answer in various dimensions, for example, a target answer in a text and video combination dimension.

[0091]The target answer may correspond to a plurality of candidate answers, and the additional question features corresponding to the target answer include an additional question feature corresponding to each candidate answer corresponding to the target answer. For example, when the target answer is “mmm”, the target answer corresponds to the candidate answer e1, the candidate answer e2, the candidate answer e3, and the candidate answer e4, and therefore, the additional question features corresponding to the target answer may include: an additional question feature corresponding to the candidate answer e1 (the candidate answer e1 is obtained based on the initial question feature and the target question d1 corresponding to the additional question feature c1, such that the candidate answer e1 corresponds to the additional question feature c1), an additional question feature corresponding to the candidate answer e2, an additional question feature corresponding to the candidate answer e3, and an additional question feature corresponding to the candidate answer e4.

[0092]So far, the target answer of the original input question is successfully output, and data processing based on the multi-modal model is completed.

[0093]In a possible embodiment, in order to enable a model to have a self-reasoning capability, a complex question may be decomposed into a simpler multi-step program, and method verification is performed in each step in multiple manners. For example, as shown in FIG. 3, FIG. 3 is a schematic diagram of a self-reasoning method based on a multi-modal model, and the method may include processes, such as a question input process, a question template generation process, a question logic analysis process, a multi-modal model (that is, multi-modal large model) process, and a technical verification process.

[0094]First, in the question input process, an original input question may be acquired, and an

example of the original input question may be “help me search for pedestrians appearing at an intersection of Chuangxin Road at about 3:00 pm on March 21, 2022, and of these pedestrians, who is most likely to be a target of the behavior X occurring at the intersection at this time point”.

[0095]Second, in the question template generation process, an initial question feature matching a question template may be determined based on the original input question. For example, the original input question is converted in a mathematical logic form, and a key value is replaced with a variable through key-value mapping to obtain the initial question feature and a modified question. For example, the original input question is modified to the following question: help me search for pedestrians appearing at B at about A, and of these pedestrians, who is most likely to be C, where A represents 3:00 pm on March 21, 2022; B represents intersection of Chuangxin Road; C represents behavior X.

[0096]Third, in the question logic analysis process, a plurality of additional question features associated with the initial question feature may be acquired, and a plurality of target questions are generated based on the initial question feature and the plurality of additional question features.

[0097]For example, based on multiple verification and cross-check methods, analysis solutions are generated by using different approaches, a prompt is provided to the multi-modal target model, additional context is generated for a question, and a logically modified question may be: a pedestrian appearing at A and B and having the behavior C may have a feature K, where K is changeable, that is, K indicates an additional question feature, such that the plurality of target questions may be generated by changing the additional question feature K, and each target question may include the initial question feature and the additional question features.

[0098]Fourth, in the multi-modal model (that is, multi-modal large model) process, each target question may be input into a multi-modal target model, and candidate answers corresponding to each target question is output by the multi-modal target model. For example, the multi-modal target model is used to analyze a logically modified question (that is, the target question) to output candidate answers, where the candidate answers use a capability of the multi-modal target model, that is, the multi-modal target model analyzes a possibility that the target question includes various dimensions such as an image, a video, a text, and a voice, the multi-modal target model compares output results to find a consensus from information in various dimensions, and the multi-modal target model may provide a higher confidence level, such that the candidate answers corresponding to the target question are obtained, that is, the candidate answers are correct and reliable.

[0099]For example, the multi-modal target model automatically analyzes pedestrians having the behavior X and conforming to a time and a place, and meanwhile, determines identity information of the pedestrians based on biological features, performs automatic reasoning and analysis by using video structured features, such as a dress color, and whether carrying a backpack, plus big data analysis to obtain a most likely target object, and quickly locates a current possible position of the target object by using space position information.

[0100]Fifth, in the technical verification process, a target answer corresponding to the original input question may be determined based on the candidate answers corresponding to each target question. For example, in order to ensure that a consensus is reached in outputs of various expressions (that is, the plurality of target questions), the candidate answers corresponding to each target question may be determined, and a candidate answer with the most frequently occurring is determined as the target answer, ensuring final output accuracy. In the end, a highest confidence list is output ceaselessly and circularly, and a most likely target object is output finally, then an additional question feature corresponding to the output target object is determined, such that the whole self-reasoning process is completed, and the target answer corresponding to the original input question is obtained.

[0101]As can be seen from the above technical solutions, in the embodiments of the present application, the target answer is obtained with the help of the multi-modal target model. Moreover, an accurate and reliable target answer is obtained through the plurality of target questions, and the target answer corresponding to the original input question can be searched for from an enormous volume of data only by consuming a small amount of computing resources, in which way, time for acquiring the target answer is shorter, saving computing resources of a server and time consumption, quickly locating the target answer corresponding to the original input question, providing abnormality logic analysis and reasoning capabilities, implementing automatic abnormality diagnosis, giving the accurate and reliable target answer through an analysis capability of the multi-modal target model, and comprehensively considering possibilities of various dimensions such as an image, a video and a voice. Through capabilities of the multi-modal target model, the user only needs to input simple descriptive characters, such that the multi-modal target model can perform automatic reasoning and quickly locate the target object. Performance of the multi-modal target model on a logic reasoning problem can be improved, and the multi-modal target model can perform self-reasoning to complete optimization requirements. In the embodiments of the present application, abnormality problems can be quickly located, and meanwhile, abnormality logic analysis and reasoning is given, implementing automatic abnormality diagnosis.

[0102]Based on the same application concept as the above method, an embodiment of the present application provides a data processing apparatus based on a multi-modal model. As shown in FIG. 4, FIG. 4 is a schematic structure diagram of the apparatus. The apparatus may include:

[0103]a determining module 41, configured to determine an initial question feature matching a question template based on an original input question; an acquiring module 42, configured to acquire a plurality of additional question features associated with the initial question feature; a generating module 43, configured to generate a plurality of target questions based on the initial question feature and the plurality of additional question features, each target question including the initial question feature and at least one additional question feature, where the acquiring module 42 is further configured to, for each target question, input the target question into a trained multi-modal target model, and output, by the multi-modal target model, candidate answers corresponding to the target question; the determining module 41 is further configured to determine a target answer corresponding to the original input question based on the candidate answers corresponding to each target question, the target answer being an output result of the original input question.

[0104]In some embodiments, the determining module 41, when determining the initial question feature matching the question template based on the original input question, is specifically configured to: if the question template includes a feature type, determine whether a key value matching the feature type exists in the original input question; if the key value matching the feature type exists, determine the initial question feature based on the key value; if the key value matching the feature type does not exist, analyze the original input question to obtain an initial question feature matching the feature type; or, send prompt information for the feature type to a user, provide, by the user, a new original input question based on the prompt information, and if the key value matching the feature type exists in the new original input question, determine the initial question feature based on the key value.

[0105]In some embodiments, the acquiring module 42, when acquiring the plurality of additional question features associated with the initial question feature, is specifically configured to: input the initial question feature into a trained feature analysis target model, and output, by the feature analysis target model, the plurality of additional question features associated with the initial question feature; or input the initial question feature into the multi-modal target model, and output, by the multi-modal target model, the plurality of additional question features associated with the initial question feature; or analyze based on the initial question feature to obtain the plurality of additional question features associated with the initial

question feature.

[0106]In some embodiments, the feature analysis target model is a feature analysis target model matching a target scenario type; the question template is a question template matching the target scenario type, where the target scenario type includes a behavior abnormality type, a vehicle abnormality type, or an industry abnormality type.

[0107]In some embodiments, the generating module 43, when generating the plurality of target questions based on the initial question feature and the plurality of additional question features, is specifically configured to: select a candidate additional question feature from the plurality of additional question features; after the initial question feature and the candidate additional question feature are combined, for each combination, add context information to the initial question feature and/or the candidate additional question feature to obtain a target question corresponding to the candidate additional question feature.

[0108]In some embodiments, the determining module 41, when determining the target answer corresponding to the original input question based on the candidate answers corresponding to each target question, is specifically configured to: filter candidate answers with repeated contents from all candidate answers based on the candidate answers corresponding to each target question to obtain at least one remaining candidate answer, and determine a number of occurrences of each remaining candidate answer in all the candidate answers; determine a remaining candidate answer with a maximum number of occurrences as the target answer corresponding to the original input question; or, if a ratio value of occurrences of a remaining candidate answer corresponding to a maximum number of occurrences in all the candidate answers is greater than a preset threshold, determine the remaining candidate answer with the maximum number of occurrences as the target answer corresponding to the original input question.

[0109]In some embodiments, the determining module 41, after determining the target answer corresponding to the original input question based on the candidate answers corresponding to each target question, is further configured to: output target data corresponding to the target answer, the target data including the target answer, or the target data including the target answer and additional question features corresponding to the target answer, where the target answer includes an answer of at least one dimension of an image, a text, a sound, or a video; the target answer corresponds to a plurality of candidate answers, and the additional question features corresponding to the target answer include an additional question feature corresponding to each candidate answer corresponding to the target answer.

[0110]Based on the same application concept as the above method, an embodiment of the

present application provides an electronic device. As shown in FIG. 5, the electronic device includes: a processor 51, and a machine readable storage medium 52 storing machine executable instructions (that is, data processing instructions based on a multi-modal model) that can be executed by the processor 51, where the processor 51 is configured to execute the machine executable instructions to implement the data processing method based on the multi-modal model disclosed in the above embodiments of the present application.

[0111]Based on the same application concept as the above method, an embodiment of the present application provides a machine readable storage medium having several computer instructions stored thereon, where the computer instructions are executed by a processor to implement the data processing method based on the multi-modal model disclosed in the above embodiments of the present application.

[0112]The machine readable storage medium may be any electronic, magnetic, optical, or other physical storage apparatus, and may contain or store information, such as executable instructions, and data. For example, the machine readable storage medium may be: a RAM (Random Access Memory), a volatile memory, a non-volatile memory, a flash memory, a storage drive (such as hard drive), a solid state hard disk, any type of storage disk (such as optical discs, and DVDs), or similar storage medium, or a combination of them.

[0113]The system, apparatus, module or unit set forth in the above examples may be specifically implemented by a computer entity, or implemented by a product having a certain function. A typical implementation device is a computer, and a specific form of the computer may be a personal computer, a laptop computer, a cellular phone, a camera phone, a smart phone, a personal digital assistant, a media player, a navigation device, an e-mail transceiver device, a game console, a tablet computer, a wearable device, or a combination of any of these devices.

[0114]For the convenience of description, during the description of the apparatus, the apparatus is divided into various units according to their functions, which are described respectively. Of course, when the present application is implemented, functions of each unit may be realized in one or more software and/or hardware.

[0115]Those skilled in the art should understand that the embodiments of the present application may be provided as methods, systems, or computer program products. Accordingly, the present application may take the form of a pure hardware embodiment, a pure software embodiment, or an embodiment combining software and hardware aspects. Furthermore, the embodiments of the present application may take the form of a computer program product implemented on one or more computer usable storage media (including, but not limited to, disk

storage, CD-ROM, optical storage, etc.) having computer usable program codes therein.

[0116]The present application is described with reference to flowcharts and/or block diagrams illustrating methods, apparatuses (systems), and computer program products according to the embodiments of the present application. It should be understood that each process and/or block, and combinations of processes and/or blocks in the flowcharts and/or block diagrams can be implemented by computer program instructions. These computer program instructions may be provided to a processor of a general-purpose computer, a special-purpose computer, an embedded processing machine or other programmable data processing device to produce a machine, so that the instructions executed by the processor of the computer or other programmable data processing device produce an apparatus for realizing functions specified in one or more processes in a flowchart and/or one or more blocks in a block diagram.

[0117]Furthermore, these computer program instructions may be stored in a computer readable memory capable of directing a computer or other programmable data processing device to operate in a particular manner, so that the instructions stored in the computer readable memory produce a product including an instruction device. The instruction device can realize functions specified in one or more processes in a flowchart and/or one or more blocks in a block diagram.

[0118]These computer program instructions may be loaded on a computer or other programmable data processing device, so that a series of operation steps are performed on the computer or other programmable device to produce processing implemented by computer, and the instructions executed on the computer or other programmable device provide steps for realizing functions specified in one or more processes in a flowchart and/or one or more blocks in a block diagram.

[0119]The above contents are some embodiments of the present application, which are not intended to limit the present application. For those skilled in the art, various modifications and variations can be made to the application. Any modification, equivalent replacement, improvement, etc. made within the spirit and principle of the present application shall fall into the scope of the claims of the present application.

CLAIMS

1. A data processing method based on a multi-modal model, comprising:

determining an initial question feature matching a configured question template based on an original input question;

acquiring a plurality of additional question features associated with the initial question feature based on the initial question feature;

generating a plurality of target questions based on the initial question feature and the plurality of additional question features, each of the plurality of target questions comprising the initial question feature and at least one of the plurality of additional question features;

for each of the plurality of target questions, inputting the target question into a trained multi-modal target model, and outputting, by the multi-modal target model, candidate answers corresponding to the target question; and

determining a target answer corresponding to the original input question based on the candidate answers corresponding to each of the plurality of target questions.

2. The method according to claim 1, wherein determining the initial question feature matching the configured question template based on the original input question comprises:

if the question template comprises a feature type, determining whether a key value matching the feature type exists in the original input question;

if the key value matching the feature type exists, determining the initial question feature based on the key value; and

if the key value matching the feature type does not exist, analyzing the original input question to obtain an initial question feature matching the feature type; or, sending prompt information for the feature type to a user, providing, by the user, a new original input question based on the prompt information, and if the key value matching the feature type exists in the new original input question, determining the initial question feature based on the key value.

3. The method according to claim 1, wherein acquiring the plurality of additional question features associated with the initial question feature based on the initial question feature comprises:

inputting the initial question feature into a trained feature analysis target model, and outputting, by the feature analysis target model, the plurality of additional question features associated with the initial question feature; or

inputting the initial question feature into the multi-modal target model, and outputting, by the multi-modal target model, the plurality of additional question features associated with the initial question feature; or

analyzing based on the initial question feature to obtain the plurality of additional question features associated with the initial question feature.

4. The method according to claim 3, wherein

the feature analysis target model is a feature analysis target model matching a target scenario type; and

the question template is a question template matching the target scenario type,

wherein the target scenario type comprises a behavior abnormality type, a vehicle abnormality type, or an industry abnormality type.

5. The method according to claim 1, wherein generating the plurality of target questions based on the initial question feature and the plurality of additional question features comprises:

selecting a candidate additional question feature from the plurality of additional question features; and

after the initial question feature and the candidate additional question feature are combined, for each combination, adding context information to the initial question feature and/or the candidate additional question feature to obtain a target question corresponding to the candidate additional question feature.

6. The method according to claim 1, wherein determining the target answer corresponding to the original input question based on the candidate answers corresponding to each of the plurality of target questions comprises:

filtering candidate answers with repeated contents from all candidate answers based on the candidate answers corresponding to each of the plurality of target questions to obtain at least one remaining candidate answer, and determining a number of occurrences of each of the at least one remaining candidate answer in all the candidate answers; and

determining a remaining candidate answer with a maximum number of occurrences as the target answer corresponding to the original input question; or, if a ratio value of occurrences of a remaining candidate answer corresponding to a maximum number of occurrences in all the candidate answers is greater than a preset threshold, determining the remaining candidate answer with the maximum number of occurrences as the target answer corresponding to the

original input question.

7. The method according to claim 6, wherein, after determining the target answer corresponding to the original input question based on the candidate answers corresponding to each of the plurality of target questions, the method further comprises:

outputting target data corresponding to the target answer, the target data comprising the target answer, or the target data comprising the target answer and additional question features corresponding to the target answer;

wherein the target answer comprises an answer of at least one dimension of an image, a text, a sound, or a video; and

wherein the target answer corresponds to a plurality of candidate answers, and the additional question features corresponding to the target answer comprise an additional question feature corresponding to each candidate answer corresponding to the target answer.

8. A data processing apparatus based on a multi-modal model, comprising:

a determining module, configured to determine an initial question feature matching a question template based on an original input question;

an acquiring module, configured to acquire a plurality of additional question features associated with the initial question feature; and

a generating module, configured to generate a plurality of target questions based on the initial question feature and the plurality of additional question features, each of the plurality of target questions comprising the initial question feature and at least one of the plurality of additional question features,

wherein the acquiring module is further configured to, for each of the plurality of target questions, input the target question into a trained multi-modal target model, and output, by the multi-modal target model, candidate answers corresponding to the target question; and

the determining module is further configured to determine a target answer corresponding to the original input question based on the candidate answers corresponding to each of the plurality of target questions, the target answer being an output result of the original input question.

9. The apparatus according to claim 8, wherein

the determining module, when determining the initial question feature matching the question template based on the original input question, is configured to: if the question template

comprises a feature type, determine whether a key value matching the feature type exists in the original input question; if the key value matching the feature type exists, determine the initial question feature based on the key value; if the key value matching the feature type does not exist, analyze the original input question to obtain an initial question feature matching the feature type; or, send prompt information for the feature type to a user, provide, by the user, a new original input question based on the prompt information, and if the key value matching the feature type exists in the new original input question, determine the initial question feature based on the key value;

the acquiring module, when acquiring the plurality of additional question features associated with the initial question feature, is configured to: input the initial question feature into a trained feature analysis target model, and output, by the feature analysis target model, the plurality of additional question features associated with the initial question feature; or input the initial question feature into the multi-modal target model, and output, by the multi-modal target model, the plurality of additional question features associated with the initial question feature; or analyze based on the initial question feature to obtain the plurality of additional question features associated with the initial question feature;

the feature analysis target model is a feature analysis target model matching a target scenario type; the question template is a question template matching the target scenario type, wherein the target scenario type comprises a behavior abnormality type, a vehicle abnormality type, or an industry abnormality type;

the generating module, when generating the plurality of target questions based on the initial question feature and the plurality of additional question features, is configured to: select a candidate additional question feature from the plurality of additional question features; after the initial question feature and the candidate additional question feature are combined, for each combination, add context information to the initial question feature and/or the candidate additional question feature to obtain a target question corresponding to the candidate additional question feature;

the determining module, when determining the target answer corresponding to the original input question based on the candidate answers corresponding to each of the plurality of target questions, is configured to: filter candidate answers with repeated contents from all candidate answers based on the candidate answers corresponding to each of the plurality of target questions to obtain at least one remaining candidate answer, and determine a number of occurrences of each of the at least one remaining candidate answer in all the candidate answers; determine a remaining candidate answer with a maximum number of occurrences as the target

answer corresponding to the original input question; or, if a ratio value of occurrences of a remaining candidate answer corresponding to a maximum number of occurrences in all the candidate answers is greater than a preset threshold, determine the remaining candidate answer with the maximum number of occurrences as the target answer corresponding to the original input question; and

the determining module, after determining the target answer corresponding to the original input question based on the candidate answers corresponding to each of the plurality of target questions, is further configured to: output target data corresponding to the target answer, the target data comprising the target answer, or the target data comprising the target answer and additional question features corresponding to the target answer, wherein the target answer comprises an answer of at least one dimension of an image, a text, a sound, or a video; the target answer corresponds to a plurality of candidate answers, and the additional question features corresponding to the target answer comprise an additional question feature corresponding to each candidate answer corresponding to the target answer.

10. An electronic device, comprising: a processor, and a machine readable storage medium storing machine executable instructions that can be executed by the processor, wherein the processor is configured to execute the machine executable instructions to implement the method according to any one of claims 1-7.

11. A machine readable storage medium having several computer instructions stored thereon, wherein the computer instructions are executed by a processor to implement the method according to any one of claims 1-7.