

IMAGE PROCESSING AND MODEL TRAINING METHODS AND APPARATUSES, DEVICE AND STORAGE MEDIUM

TECHNICAL FIELD

- [0001] The application relates to, but is not limited to, the technical field of information, and in particular to image processing and model training methods and apparatuses, a device, a storage medium and a computer program product.

BACKGROUND

- [0002] In the field of image processing, a training sample set is usually used to train a model. In the related art, since the training sample set usually only contains image samples of some image categories, the trained model only can recognize the image categories learned in the training sample set. For the image categories that do not appear in the training sample set, the recognition accuracy is relatively low usually, resulting in poor generalization ability of the model.

SUMMARY

- [0003] In view of this, embodiments of the application provides image processing and model training methods and apparatuses, a device, a storage medium and a computer program product.
- [0004] The technical solution of the embodiments of the application is implemented as follow.
- [0005] In on aspect, an embodiment of the application provides an image processing method, which includes the following operations.
- [0006] An image feature of an image to be processed is obtained.
- [0007] Category prompt information of each image category in an image category set is determined based on the image feature.
- [0008] A text feature of each image category is determined based on each category prompt information.
- [0009] Classified processing is performed on the image feature based on the text feature of each image category to obtain a classification result of the image to be processed.
- [0010] In another aspect, an embodiment of the application provides a model training method, which includes the following operations.
- [0011] An image sample is processed by using a model to be trained to obtain a classification result of the image sample. Processing the image sample by using the model includes: an image feature of the image sample is obtained; category prompt information of each image category in an image category set is determined based on the image feature; a text feature of each image category is determined based on each category prompt information; classified processing is performed on the image feature

based on the text feature of each image category to obtain a classification result of the image sample.

- [0012] At least one updating is performed on network parameter of the model based on the classification result and a category label of the image sample to obtain the trained model.
- [0013] In another aspect, an embodiment of the application provides an image processing apparatus, which includes an obtaining module, a first determination module, a second determination module and a classification module.
- [0014] The obtaining module is configured to obtain an image feature of an image to be processed.
- [0015] The first determination module is configured to determine category prompt information of each image category in an image category set based on the image feature.
- [0016] The second determination module is configured to determine a text feature of each image category based on each category prompt information.
- [0017] The classification module is configured to perform classified processing on the image feature based on the text feature of each image category to obtain a classification result of the image to be processed.
- [0018] In yet another aspect, an embodiment of the application provides a model training apparatus, which may include a processing module and an updating module.
- [0019] The processing module is configured to process an image sample by using a model to be trained to obtain a classification result of the image sample. Processing the image sample by using the model includes: an image feature of the image sample is obtained; category prompt information of each image category in an image category set is determined based on the image feature; a text feature of each image category is determined based on each category prompt information; and classified processing is performed on the image feature based on the text feature of each image category to obtain a classification result of the image sample.
- [0020] The updating module is configured to perform at least one updating on a network parameter of the model based on the classification result and a category label of the image sample to obtain the trained model.
- [0021] In yet another aspect, an embodiment of the application provides a computer device, which includes a memory and a processor. The memory stores a computer program capable of running on the processor. When executing the program, the processor implements part or all steps of the above method.
- [0022] In still another aspect, an embodiment of the application provides a computer-readable storage medium, in which a computer program is stored. When executed by a processor, the computer program implements part or all steps of the above method.
- [0023] In still another aspect, an embodiment of the application provides a computer program, which includes a computer-readable code. When the computer-readable code runs in a computer device, a processor in a display device implements part or all steps in the above method.
- [0024] In still another aspect, an embodiment of the application provides a computer

program product, which includes a non-transitory computer-readable storage medium, in which a computer program is stored. When read and executed by a computer, the computer program implements part or all steps in the above method.

- [0025] In the embodiments of the application, an image feature of an image to be processed is obtained, category prompt information of each image category in the image category set is determined based on the image feature, a text feature of each image category is determined based on each category prompt information, and classified processing is performed on the image feature based on the text feature of each image category to obtain a classification result of the image to be processed. In this way, for each image to be processed, since the category prompt information of each image category is dynamically determined based on the image feature of the image to be processed, the text feature of each image category obtained based on the dynamically determined category prompt information can better reflect the feature of the image to be processed, so as to enhance the accuracy of image recognition. In addition, since the static category prompt information in the related art is usually set by those skilled in the art according to apriori information such as experience and an actual application scene, or is determined after learning through model training, the static category prompt information may over-fit the apriori information and/or the image category learned in the model training process. Therefore, compared with the static category prompt information, the dynamically determined category prompt information in the embodiments of the application can also reduce the over-fitting of the apriori information and/or the image categories learned in the model training process. Thus, based on the dynamically determined category prompt information, the ability of recognizing image categories which are not considered in the apriori information and/or image categories which are not involved in the model training process can be improved, and then the generalization ability of image classification and recognition can be improved.

BRIEF DESCRIPTION OF THE DRAWINGS

- [0026] FIG. 1 illustrates an implementation flowchart of an image processing method provided by an embodiment of the application.
- [0027] FIG. 2 illustrates an implementation flowchart of an image processing method provided by an embodiment of the application.
- [0028] FIG. 3 illustrates an implementation flowchart of an image processing method provided by an embodiment of the application.
- [0029] FIG. 4 illustrates an implementation flowchart of an image processing method provided by an embodiment of the application.
- [0030] FIG. 5A illustrates an implementation architecture diagram of a visual language model provided by an embodiment of the application.
- [0031] FIG. 5B illustrates an implementation flowchart of a model training method provided by an embodiment of the application.
- [0032] FIG. 6 illustrates a composition structure diagram of an image processing apparatus provided by an embodiment of the application.
- [0033] FIG. 7 illustrates a composition structure diagram of a model training

apparatus provided by an embodiment of the application.

- [0034] FIG. 8 illustrates a schematic diagram of a hardware entity of a computer device provided by an embodiment of the application.

DETAILED DESCRIPTION

- [0035] For making the purposes, technical solutions and advantages of the application clearer, the technical solutions of the application will further be described below in combination with the drawings and embodiments in detail. The described embodiments should not be considered as limits to the application. All other embodiments obtained by those of ordinary skill in the art without creative work shall fall within the scope of protection of the application.
- [0036] "Some embodiments" involved in the following descriptions describes a subset of all possible embodiments. However, it can be understood that "some embodiments" may be the same subset or different subsets of all the possible embodiments, and may be combined without conflicts. Term "first/second/third" involved is only for distinguishing similar objects and does not represent a specific sequence of the objects. It can be understood that "first/second/third" may be interchanged to specific sequences or orders if allowed to implement the embodiments of the application described herein in sequences except the illustrated or described ones.
- [0037] Unless otherwise defined, all technical and scientific terms in the specification have the same meaning as those skilled in the art, belonging to the application, usually understand. Terms used in the specification are only used for describing the purpose of the application, but not intended to limit the application.
- [0038] For better understanding of the embodiment of the application, an image classification and recognition scheme in the related art is described first.
- [0039] Taking a visual language model as an example, the core of the visual language model is to align the output of an image encoder and a text encoder by using a comparative learning loss function, and generate the classification weight by inputting prompt information containing a category name into the text encoder. In the process of deploying the visual language model to a downstream image recognition task, the biggest technical difficulty is the design of prompt information, and that is because designing appropriate prompt information requires a lot of time to fine tune words (such as words, Chinese characters and words) on a set of test data.
- [0040] To solve this problem, a continuous prompt information learning method is proposed in the related art. Its main idea is to transform context words in the prompt information into random vectors, and then learn by returning the gradient from the text encoder to these random vectors, so as to obtain the trained visual language model. However, in the process of implementing the embodiments of the application, the inventor found that the prompt information learned by the continuous prompt information learning method is fixed after training, that is, in the process of image recognition using the trained visual language model, the prompt information remains static, and the static prompt information is usually fitted to apriori information such as the experience of those skilled in the art and an actual application scene and/or image categories learned during model training, and is not applicable to image categories that have not appeared in a

training sample set, resulting in relatively low generalization ability of the visual language model after training.

- [0041] An embodiment of the application provides an image processing method, which may be executed by a processor of a computer device. The computer device refers to a server, a notebook computer, a tablet computer, a desktop computer, a smart TV, a set-top box, a mobile device (such as a mobile phone, a portable video player, a personal digital assistant, a special messaging device and a portable game device) and other devices with data processing capabilities. FIG. 1 illustrates an implementation flowchart of an image processing method provided by an embodiment of the application. As shown in FIG. 1, the method includes S101 to S104 below.
- [0042] At S101, an image feature of an image to be processed is obtained.
- [0043] Here, the image to be processed may be any appropriate image. During implementation, those skilled in the art may select an appropriate image to be processed according to an actual application scene, which is not limited in the embodiment of the application. For example, the image to be processed may include, but is not limited to, at least one of: an image of images, a video or an image frame sequence taken by a camera, a medical image, an image generated by a computer technology, etc.
- [0044] The image feature of the image to be processed may be extracted from the image to be processed through any appropriate feature extraction mode. For example, the image to be processed may be extracted by adopting a convolutional neural network, a recurrent neural network or a transformer-based feature extraction network to obtain the image features. During implementation, the image feature extracted in advance from the image to be processed may be obtained, and the image feature may also be obtained in real time by feature extraction of the image to be processed, which is not limited here.
- [0045] At S102, category prompt information of each image category in an image category set is determined based on the image feature.
- [0046] Here, the image category may be any appropriate category determined according to an actual service scene, which includes, but is not limited to, at least one of the category (such as human, cat, dog, rabbit and tiger) of an object contained in the image, the category (such as indoor space, airport, station and shopping mall) of a scene represented by the image, the state category (such as defective and ineffective) of a target object in the image, the posture category (such as stand, sit and lie) of a human body in the image and the like.
- [0047] The image category set may be a set of at least one image category to be recognized. During implementation, those skilled in the art may determine the image category in the image category set in an appropriate way according to an actual situation, which is not limited by the embodiment of the application. For example, the image category set may be preset, or may be determined based on a current image recognition task, or may be determined based on the source of the image to be processed.
- [0048] The category prompt information of the image category may be prompt information for describing the image category, may include description information in the form of a text, and may also include at least one word embedding vector, which is not limited here. The category prompt information of each image category in the image category set is determined based on the image feature of the image to be processed, and

the category prompt information that may reflect the feature of the image to be processed may be obtained. During implementation, those skilled in the art may determine the category prompt information of each image category in the image category set in an appropriate way based on the image feature of the image to be processed according to an actual situation, which is not limited here.

[0049] For example, when the category prompt information includes at least one word embedding vector, the image feature may be mapped to the preset dimension to obtain at least one word embedding vector with the preset dimension, and for each image category, the at least one word embedding vector with the preset dimension may be combined with the word embedding vector corresponding to the category name of the image category to obtain a combined word embedding vector set, and the word embedding vector set is the category prompt information of the image category. When the category prompt information includes text description information, the image feature may be mapped to the preset dimension to obtain at least one word embedding vector with the preset dimension, and for each image category, a word corresponding to each word embedding vector and the category name of the image category may be combined to obtain a piece of text description information, and the text description information is the category prompt information of the image category.

[0050] At S103, a text feature of each image category is determined based on each piece of category prompt information.

[0051] Here, for each image category, the text feature of the image category may be determined based on the category prompt information of the image category. During implementation, the text feature may be extracted from the category prompt information by adopting any appropriate text feature extraction algorithm, which is not limited here. For example, the category prompt information may be encoded by adopting a text encoder network, such as a transformer-based encoder network, to obtain the text feature.

[0052] At S104, classified processing is performed on the image feature based on the text feature of each image category to obtain a classification result of the image to be processed.

[0053] Here, by matching the image feature with the text feature of each image category, the matching degree between the image feature and each text feature may be determined, so as to obtain the classification result of the image to be processed.

[0054] The classification result of the image to be processed may include any appropriate information for representing the category to which the image to be processed belongs, which is not limited here. For example, the classification result of the image to be processed may include an image category to which the image to be processed belongs. For another example, the classification result of the image to be processed may include the probability that the image to be processed belongs to each image category.

[0055] In some implementation modes, the classification result of the image to be processed may be determined based on the similarity between the image feature and the text feature of each image category.

[0056] In some embodiments, the classification result includes the probability that the image to be processed belongs to each image category. The S104 may include S111 to S112 below.

[0057] At S111, the similarity between the image feature and the text feature of each image category is determined.

[0058] Here, the similarity between the image feature and the text feature of each image category may be determined in any appropriate way, which is not limited by the embodiment of the application. For example, the similarity between the image feature and the text feature may be determined based on Euclidean distance, Pearson correlation coefficient, cosine similarity, Manhattan distance and the like.

[0059] At S112, the probability that the image to be processed belongs to each image category is determined based on each similarity.

[0060] Here, according to the similarity between the image feature and the text feature of each image category, the matching degree between the image features and each text feature may be determined, so as to determine the probability that the image to be processed belongs to each image category. During implementation, those skilled in the art may determine the probability that the image to be processed belongs to each image category based on the similarity between the image feature and the text feature of each image category by adopting an appropriate way according to an actual situation, which is not limited by the embodiment of the application.

[0061] In some implementation modes, the similarity between the image feature and the text feature of each image category may be normalized to obtain the probability that the image to be processed belongs to each image category. For example, the probability $p(y|x)$ that the image to be processed belongs to the y-th image category is determined by using a normalized exponential function (softmax) in a way as shown in the formula 1 below:

$$p(y|x) = \frac{\exp(\text{sim}(x, g(t_y))/\tau)}{\sum_{i=1}^K \exp(\text{sim}(x, g(t_i))/\tau)} \quad (1)$$

where x is an image feature of the image to be processed, t_i is category prompt information of the i-th image category, $g(t_i)$ is a text feature of the i-th image category, $\text{sim}(x, g(t_i))$ is the cosine similarity between the image feature of the image to be processed and the text feature of the i-th image category, y is an integer greater than or equal to 1 and is less than or equal to K , K is an integer greater than 1, and τ is a set temperature parameter.

[0062] In the embodiment of the application, an image feature of an image to be processed is obtained, category prompt information of each image category in the image category set is determined based on the image feature, a text feature of each image category is determined based on each piece of category prompt information, and classified processing is performed on the image feature based on the text feature of each image category to obtain a classification result of the image to be processed. In this way, for each image to be processed, since the category prompt information of each image category is dynamically determined based on the image feature of the image to be processed, the text feature of each image category obtained based on the dynamically determined category prompt information can better reflect the feature of the image to be processed, so as to enhance the accuracy of image recognition. In addition, since the

static category prompt information in the related art is usually set by those skilled in the art according to apriori information such as experience and an actual application scene, or is determined after learning through model training, the static category prompt information may over-fit the apriori information and/or the image category learned in the model training process. Therefore, compared with the static category prompt information, the dynamically determined category prompt information in the embodiment of the application can also reduce the over-fitting of the apriori information and/or the image categories learned in the model training process. Thus, based on the dynamically determined category prompt information, the ability of recognizing image categories which are not considered in the apriori information and/or image categories which are not involved in the model training process can be improved, and then the generalization ability of image classification and recognition can be improved.

- [0063] An embodiment of the application provides an image processing method, which may be executed by a processor of a computer device. As shown in FIG. 2, the method includes S201 to S205 below.
- [0064] At S201, an image feature of an image to be processed is obtained.
- [0065] Here, S201 corresponds to the above S101, which may be implemented with reference to the specific implementation mode of the above S101.
- [0066] At S202, image mark information is determined based on the image feature.
- [0067] Here, the image mark information may be used to represent the image feature and is used as the mark of the image to be processed. For example, the image mark information may include a representation vector of the image feature. During implementation, those skilled in the art may determine the image mark information based on the image feature in an appropriate way according to an actual situation, which is not limited here.
- [0068] In some implementation modes, an appropriate mapping strategy may be set in advance according to an actual situation, and the image feature may be mapped based on the mapping strategy to obtain the image mark information used to represent the image feature. For example, the image feature may be mapped to the set dimension by linear mapping to obtain a mapped representation vector, which may be used as image mark information.
- [0069] In some implementation modes, a trained neural network may be used to process the image feature to obtain the image mark information used to represent the image feature.
- [0070] At S203, category prompt information of each image category in an image category set is determined based on the image mark information and set contextual semantic information.
- [0071] Here, the category name of the image category may be extended based on the image mark information and the contextual semantic information to obtain the prompt information for describing the image category, namely the category prompt information of the image category. The contextual semantic information and the image mark information may be reflected in the category prompt information of each image category. During implementation, those skilled in the art may determine the category prompt information of each image category in the image category set in an appropriate way

based on the image mark information and the set contextual semantic information according to an actual situation, which is not limited here.

- [0072] In some implementation modes, for each image category in the image category set, the image mark information, the contextual semantic information and the category name of the image category may be combined to obtain the category prompt information of the image category.
- [0073] The set contextual semantic information may be randomly initialized, may be obtained from a preset word dictionary, or may be obtained through training with image samples, which is not limited here.
- [0074] At S204, a text feature of each image category is determined based on each piece of category prompt information.
- [0075] At S205, classified processing is performed on the image feature based on the text feature of each image category to obtain a classification result of the image to be processed.
- [0076] Here, S204 to S205 correspond to the above S103 to S104. During implementation, it may refer to the specific implementation modes of the above S103 to S104.
- [0077] In the embodiment of the application, the image mark information is determined based on the image feature, and the category prompt information of each image category in the image category set is determined based on the image mark information and the contextual semantic information. In this way, the contextual semantic information and the image feature of the image to be processed are comprehensively considered in the category prompt information of each image category. Compared with the case of considering the image feature alone, the accuracy of image recognition can be further enhanced. Compared with the case of considering the contextual semantic information alone, it can reduce the problem of over-fitting the category prompt information to the image category applicable to the contextual semantic information, so as to further improve the ability of recognizing the apriori information which is dependent on the contextual semantic information and/or the image category which is not involved in the contextual semantic information learning process, and then further improve the generalization ability of image classification and recognition.
- [0078] In some embodiments, S203 may include the following operations.
- [0079] At S211, for each image category in the image category set, the category prompt information of the image category is determined based on the image mark information, the contextual semantic information and a word embedding vector corresponding to the image category.
- [0080] Here, each image category may correspond to a word embedding vector. During implementation, the feature of the category name of the image category may be mapped to a lower dimensional vector in any suitable way to obtain the word embedding vector corresponding to the image category, which is not limited by the embodiment of the application. For example, the category name of each image category may be mapped into a word embedding vector with a set dimension by using One-Hot encoding, a Continuous Bag-of-Word (CBOW) model, a Skip-Gram model, etc.

- [0081] In some implementation modes, the word embedding vector corresponding to the image category may be combined with the image mark information and the contextual semantic information to obtain the category prompt information of the image category. During implementation, the way of combining the word embedding vector corresponding to the image category, the image mark information and the contextual semantic information may be determined according to an actual situation, which is not limited here. For example, the category prompt information of the image category may be obtained through at least one operation such as combination and superposition of the word embedding vector corresponding to the image category, the image mark information and the contextual semantic information.
- [0082] In some embodiments, the contextual semantic information includes at least one contextual vector, the image mark information includes an image mark vector, and the category prompt information includes a category prompt vector. Determining the category prompt information of the image category based on the image mark information, the contextual semantic information and the word embedding vector corresponding to the image category in S211 may include S221 to S222 below.
- [0083] At S221, for each contextual vector, an image contextual vector is determined based on the image mark vector and the contextual vector.
- [0084] Here, the image mark vector and each contextual vector may be combined respectively to obtain at least one image contextual vector. During implementation, the image mark vector and the contextual vector may be combined in any appropriate way, including, but not limited to, at least one operation such as superposition, connection, cross multiplication and the like of the image mark vector and the contextual vector, which is not limited by the embodiment of the application.
- [0085] In some implementation modes, for each contextual vector, the image mark vector of the image to be processed may be superimposed with the contextual vector to obtain an image contextual vector.
- [0086] The number of contextual vectors included in the contextual semantic information may be set by those skilled in the art according to an actual situation, such as 4, 8 and 16. During implementation, those skilled in the art may determine an appropriate number of contextual vectors according to the difficulty of an actual image recognition task, the computing resources of a hardware device for executing a model training method and the like, which is not limited here. For example, when the difference among a plurality of image categories to be recognized in an image recognition task is relatively small, more contextual vectors may be set to enhance the accuracy of image recognition; when the difference among the plurality of image categories to be recognized in the image recognition task is relatively large, less contextual vectors may be set to reduce the number of parameters learned in the model training process, so as to enhance the model training efficiency. For another example, when the hardware device for executing the model training method has more computing resources, more contextual vectors may be set to enhance the accuracy of image recognition; when the hardware device for executing the model training method has less computing resources, less context vectors may be set to reduce the number of parameters learned in the model training process, so as to enhance the model training efficiency.
- [0087] At S222, category prompt vector of the image category is determined based on

each image contextual vector and the word embedding vector.

- [0088] Here, for each image category, each image contextual vector may be connected with the word embedding vector corresponding to the image category to obtain the category prompt vector of the image category. During implementation, the word embedding vector corresponding to the image category may be connected after each image contextual vector, the word embedding vector corresponding to the image category may also be connected before each image contextual vector, and the word embedding vector corresponding to the image category may also be connected between image contextual vectors. The embodiment of the application is not limited to this.
- [0089] In the embodiment, for each image category in the image category set, the category prompt information of the image category may be determined based on the image mark information of the image to be processed, the contextual semantic information and the word embedding vector corresponding to the image category. Therefore, appropriate category prompt information may be quickly determined for each image category in the image category set, so as to enhance the efficiency of image classification and recognition.
- [0090] In some embodiments, the contextual semantic information includes at least one of the followings: at least one randomly generated contextual vector; and at least one contextual vector obtained from a set word dictionary. Here, the contextual vector in the word dictionary may include the pre-determined word embedding vector representing the contextual semantics. Each word embedding vector representing the context semantics may correspond to manually designed contextual words or contextual words determined after fine-tuning based on test data. The contextual vector in the word dictionary may also include at least one contextual vector obtained after training based on an image sample in advance, which is not limited by the embodiment of the application.
- [0091] In the embodiment described above, when the contextual semantic information includes at least one randomly generated contextual vector, at least one contextual vector can be generated simply and quickly; when the contextual semantic information includes at least one contextual vector obtained from the set word dictionary, since the contextual vector in the word dictionary is a word embedding vector that may represent the contextual semantics, the accuracy of image recognition can be further enhanced.
- [0092] An embodiment of the application provides an image processing method, which may be executed by a processor of a computer device. As shown in FIG. 3, the method includes S301 to S304 below.
- [0093] At S301, feature extraction is performed on an image to be processed by using an image feature extraction network in a model to obtain an image feature of the image to be processed.
- [0094] At S302, category prompt information of each image category in an image category set is determined based on the image feature by using a prompt information generation network in the model.
- [0095] At S303, feature extraction is performed on each piece of category prompt information by using a text feature extraction network in the model to obtain a text feature of each image category.
- [0096] Here, the model may be a trained deep learning model, including an image

feature extraction network, a text feature extraction network and a prompt information generation network. During implementation, the image feature extraction network may be any appropriate network for image feature extraction, including, but not limited to, Residual Networks (ResNets), a Vision Transformer (ViT) network, etc. The text feature extraction network may be any appropriate network for text feature extraction, including, but not limited to, an transformer-based encoder network, a bidirectional long-term and short-term memory neural network, etc. The prompt information generation network may be any appropriate network for generating category prompt information of image categories based on image features, which may be a simple multi-layer perceptron, a more complex network structure, a feature processing module without the perceptron, etc. Those skilled in the art may select an appropriate network structure for the image feature extraction network, the text feature extraction network and the prompt information generation network according to an actual situation, which is not limited by the embodiment of the application.

- [0097] In some implementation modes, the model is a visual language model, and the image feature extraction network may be an image encoder network in the visual language model, which may be realized by a feature extraction network based on convolutional neural network, a recurrent neural network and/or a transformer. The text feature extraction network may be a text encoder network in the visual language model, and may adopt any appropriate encoder network structure, such as the transformer-based feature extraction network.
- [0098] At S304, classified processing is performed on the image feature based on the text feature of each image category to obtain a classification result of the image to be processed.
- [0099] Here, S304 corresponds to the above S104, which may be implemented with reference to the specific implementation mode of the above S104.
- [00100] In some embodiments, the prompt information generation network includes a mark extraction sub-network and an information generation sub-network. S302 may include S311 to S312 below.
- [00101] At S311, the image feature is processed by using the mark extraction sub-network to obtain image mark information.
- [00102] Here, the mark extraction network may be used to generate image mark information for the input image feature, and the image mark information may be used to represent the image feature and may be used as the mark of the image to be processed. For example, the image mark information may include a representation vector of the image features.
- [00103] During implementation, the mark extraction network may be a simple multi-layer perceptron, a more complex network structure, or a feature processing module without the perceptron. Those skilled in the art may adopt an appropriate structure to realize the mark extraction network according to an actual situation. The embodiment of the application is not limited to this. For example, the mark extraction network may be constructed by adopting a two-layer bottleneck structure, and the input dimension is reduced by 16 in a hidden layer.
- [00104] At S312, category prompt information of each image category in an image

category set is determined based on the image mark information and set contextual semantic information by using the information generation sub-network.

[00105] Here, the information generation sub-network may be any appropriate network that may process the image mark information and the set contextual semantic information to obtain the category prompt information of each image category, which is not limited here. During implementation, those skilled in the art may determine the appropriate information generation sub-network according to the way of determining the category prompt information of each image category based on the image mark information and the contextual semantic information.

[00106] In some implementation modes, the contextual semantic information may include at least one contextual vector, the image mark information may include an image mark vector, and the category prompt information may include a category prompt vector. The information generation sub-network may be used to determine at least one image contextual vector based on the image mark vector and each contextual vector. Based on each image contextual vector and a word embedding vector corresponding to each image category, the category prompt vector of each image category is determined.

[00107] In the embodiment of the application, feature extraction is performed on the image to be processed by using the image feature extraction network in the model to obtain the image feature of the image to be processed. The category prompt information of each image category in the image category set is determined based on the image feature by using the prompt information generation network in the model. Feature extraction is performed on each piece of category prompt information by using the text feature extraction network in the model to obtain the text feature of each image category. In this way, since each network in the deep learning model is learnable, the accuracy of the extracted image feature of the image to be processed and the text features of each image category can be enhanced through the deep learning model, so as to enhance the accuracy of the classification result of the image to be processed.

[00108] An embodiment of the application provides a model training method, which may be executed by a processor of a computer device. As shown in FIG. 4, the method includes S401 to S402 below.

[00109] At S401, an image sample is processed by using a model to be trained to obtain a classification result of the image sample, where processing the image sample by using the model includes: an image feature of the image sample is obtained; category prompt information of each image category in an image category set is determined based on the image feature; a text feature of each image category is determined based on each piece of category prompt information; and classified processing is performed on the image features based on the text feature of each image category to obtain a classification result of the image sample.

[00110] Here, the model to be trained may be any appropriate deep learning model, which is not limited here. During implementation, those skilled in the art may adopt an appropriate network structure to construct the model to be trained according to an actual situation.

[00111] The process of processing the image sample by using the model to be trained to obtain the classification result of the image sample corresponds to the process of processing the image to be processed in S101 to S104 in the above embodiment. During

implementation, the process may refer to the specific implementation modes of the above S101 to S104.

[00112] At S402, at least one updating is performed on a network parameter of the model based on the classification result and a category label of the image sample to obtain the trained model.

[00113] Here, whether to update the network parameter of the model may be determined based on the classification result and the category label of the image sample. When it is determined to update the network parameter of the model, the network parameter of the model may be updated by using an appropriate parameter updating algorithm, and the classification result of the image sample may be re-determined by using the model with the updated parameter, so that whether to continue to update the network parameter of the model is determined based on the re-determined classification result and the category label of the image sample. When it is determined not to continue to update the network parameter of the model, the finally updated model is determined as the trained model.

[00114] In some implementation modes, the loss value may be determined based on the classification result and the category label of the image sample, and the network parameter of the model may be updated when the loss value does not meet a preset condition. When the loss value meets a preset condition or the number of times of updating the network parameter of the model reaches a set threshold, the updating of the network parameter of the model may be stopped, and the finally updated model is determined as the trained model. The preset condition may include, but is not limited to, at least one of: the loss value is less than a set loss threshold, the change of the loss value converges, etc. During implementation, the preset condition may be set according to an actual situation, which is not limited by the embodiment of the application.

[00115] The way of updating the network parameter of the model may be determined according to an actual situation, and includes, but is not limited to, at least one of a gradient descent method, a Newton momentum method and the like, which is not limited here.

[00116] In the embodiment of the application, an image feature of an image sample is obtained by using a model to be trained, category prompt information of each image category in an image category set is determined based on the image feature, a text feature of each image category is determined based on each piece of category prompt information, classified processing is performed on the image feature based on the text feature of each image category to obtain a classification result of the image sample, and at least one updating is performed on a network parameter of the model based on the classification result and the category label of the image sample to obtain the trained model. In this way, in the model training process, since the category prompt information of each image category is dynamically determined based on the image feature of the image sample, the text feature of each image category obtained based on the dynamically determined category prompt information can better reflect the feature of the image sample, so as to enhance the image recognition accuracy of the trained model. Moreover, compared with the static category prompt information, the dynamically determined category prompt information can also reduce the over-fitting of the image categories learned in the model training process. Thus, based on the dynamically determined category prompt information, the ability of recognizing image categories which are not

involved in the model training process can be improved, and then the generalization ability of image classification and recognition can be improved. In some implementation modes, the trained model may be migrated to other cross domain image recognition tasks with the image sample used for training, and the performance of the migrated model in other cross domain image recognition tasks can be improved.

- [00117] In some embodiments, S402 may include S411 to S412 below.
- [00118] At S411, the loss value is determined based on the classification result and the category label of the image sample.
- [00119] At S412, at least one updating is performed on the network parameter of the model based on the loss value to obtain the trained model.
- [00120] Here, the loss value between the classification result of the image sample and the category label may be determined by adopting any appropriate loss function, which is not limited here. For example, the loss function may include, but is not limited to, at least one of an absolute value loss function, a least square error loss function, a cosine loss function, a cross entropy loss function, etc.
- [00121] In some embodiments, the method further includes at least one of S421 or S422 below.
- [00122] At S421, the contextual semantic information is initialized based on at least one randomly generated contextual vector.
- [00123] At S422, the contextual semantic information is initialized based on at least one contextual vector in a set word dictionary.
- [00124] Here, at least one randomly generated contextual vector may be determined as the initial value of the contextual semantic information, so that the initial value of the contextual semantic information can be determined simply and quickly, which is convenient for model training. At least one contextual vector obtained from the word dictionary may also be determined as the initial value of the contextual semantic information. In this way, since the contextual vector in the word dictionary is a word embedding vector that may represent the contextual semantics, the initial image recognition ability of the model can be improved, so as to enhance the training efficiency of the model, and further enhance the accuracy of image recognition of the trained model.
- [00125] In some embodiments, the model includes an image feature extraction network, a prompt information generation network and a text feature extraction network. S401 may include S431 to S434 below.
- [00126] At S431, feature extraction is performed on an image sample by using the image feature extraction network to obtain an image feature of the image sample.
- [00127] At S432, category prompt information of each image category in an image category set is determined based on the image feature by using the prompt information generation network.
- [00128] At S433, feature extraction is performed on each piece of category prompt information by using the text feature extraction network to obtain a text feature of each image category.

- [00129] At S434, classified processing is performed on the image feature based on the text feature of each image category to obtain the classification result of the image sample.
- [00130] Here, S431 to S434 correspond to the above S301 to S304, respectively. During implementation, it may refer to the specific implementation modes of the above S301 to S304.
- [00131] In some implementation modes, the network parameter of the image feature extraction network and the network parameter of the text feature extraction network may be pre-trained and fixed through any appropriate pre-training method. In the current training process of the model, the network parameter of the image feature extraction network and the network parameter of the text feature extraction network remain unchanged. In this way, the number of parameters to be learned in the model training process may be reduced, so as to enhance the model training efficiency.
- [00132] In some embodiments, S402 may include the following operations.
- [00133] At S441, at least one updating is performed on the network parameter of the prompt information generation network based on the classification result and the category label of the image sample to obtain the trained model.
- [00134] Here, whether to update the network parameter of the prompt information generation network may be determined based on the classification result and the category label of the image sample. When it is determined to update the network parameter of the prompt information generation network, the network parameter of the prompt information generation network may be updated by using an appropriate parameter updating algorithm, and the classification result of the image sample may be re-determined based on the prompt information generation network with the updated parameter, so that whether to continue to update the network parameter of the prompt information generation network is determined based on the re-determined classification result and the category label of the image sample. When it is determined not to continue to update the network parameter of the prompt information generation network, the network parameter of the finally updated information generation network is determined as the network parameter of the information generation network in the trained model, so as to obtain the trained model.
- [00135] In some implementation modes, the loss value may be determined based on the classification result and the category label of the image sample, and the network parameter of the information generation network may be updated when the loss value does not meet a preset condition. When the loss value meets the preset condition or the number of times of updating the network parameter of the information generation network reaches a set threshold, the updating of the network parameter of the information generation network may be stopped, and the trained model is obtained. The preset condition may include, but is not limited to, at least one of: the loss value is less than a set loss threshold, the change of the loss value converges, etc. During implementation, the preset condition may be set according to an actual situation, which is not limited by the embodiment of the application.
- [00136] The way of updating the network parameter of the prompt information generation network may be determined according to an actual situation, and includes, but not limited to, at least one of a gradient descent method, a Newton momentum method and the like, which is not limited here.

- [00137] In some implementation modes, the learnable network parameter in the prompt information generation network may include, but is not limited to, at least one of the network parameter of the mark extraction sub-network, contextual semantic information, etc.
- [00138] In the embodiment, at least one updating is performed on the network parameter of the prompt information generation network in the model to be trained based on the classification result and the category label of the image sample to obtain the trained model, so that the number of parameters to be learned in the model to be trained can be reduced, the model training can be applied to the learning with a small number of samples, so as to reduce the number of samples used in model training and enhance the training efficiency.
- [00139] In some embodiments, the prompt information generation network includes a mark extraction sub-network and an information generation sub-network. S432 may include S451 to S452 below.
- [00140] At S451, the image feature is processed by using the mark extraction sub-network to obtain image mark information.
- [00141] At S452, category prompt information of each image category in an image category set is determined based on the image mark information and contextual semantic information by using the information generation sub-network.
- [00142] Here, S451 to S452 correspond to the above S311 to S312, respectively. During implementation, it may refer to the specific implementation modes of the above S311 to S312.
- [00143] In some embodiments, S402 may include the following operations.
- [00144] At S461, at least one updating is performed on network parameters of the mark extraction sub-network and the contextual semantic information based on the classification result and the category label to obtain the trained model.
- [00145] Here, the network parameter of the mark extraction sub-network and the contextual semantic information may be updated at the same time. During implementation, the way of updating the network parameter of the mark extraction sub-network and the contextual semantic information may be determined according to an actual situation, which includes, but is not limited to, at least one of a gradient descent method, a Newton momentum method and the like, which is not limited here. The way of updating the network parameter of the mark extraction sub-network and the way of updating the contextual semantic information may be the same or different.
- [00146] In the embodiment, at least one updating is performed on the network parameter of the mark extraction sub-network and the contextual semantic information based on the classification result and the category label to obtain the trained model. In this way, the network parameter of the mark extraction sub-network in the model to be trained and the contextual semantic information may be learned at the same time, which can further reduce the number of parameter to be learned in the model to be trained, further reduce the number of samples used in model training, and enhance the training efficiency.
- [00147] An embodiment of the application provides a visual language model. FIG. 5A

illustrates an implementation architecture diagram of a visual language model provided by an embodiment of the application. As shown in FIG. 5A, the visual language model includes an image feature extraction network 51, a prompt information generation network 52 and a text feature extraction network 53. The prompt information generation network 52 includes a mark extraction sub-network 521 and an information generation sub-network 522. The image feature extraction network 51 is used for performing feature extraction on an input image g_1 to be processed to obtain the image feature x_1 of the image g_1 to be processed. The mark extraction sub-network 521 is used for processing the image feature f_1 to obtain image mark information π . The information generation sub-network 522 is used for: superposing the image mark information π with each contextual vector in a set of contextual vectors $[v_1, v_2, \dots, v_M]$ to obtain a set of image mark vectors $[v_1+\pi, v_2+\pi, \dots, v_M+\pi]$; combining the set of image mark vectors with a word embedding vector c_i of the category name of each image category in an image category set to obtain category prompt information t_i of each image category. The text feature extraction network 53 is used for: performing feature extraction on the category prompt information t_i of each image category to obtain the text feature y_i of each image category; determining the probability p_i that the image to be processed belongs to each image category based on the image feature x_1 and the text feature y_i of each image category. M is an integer greater than 1. The number of image categories in the image category set is K . K is a positive integer. i is a positive integer not greater than K .

- [00148] During implementation, the image feature extraction network 51 may be a transformer-based encoder network, or may be a convolutional neural network and the like. The text feature extraction network 53 may be a transformer-based encoder network. In the prompt information generation network 52, the mark extraction sub-network 521 may be a simple multi-layer perceptron or contain a more complex network structure. The information generation sub-network 522 may be a simple computing module. The contextual vectors used in the information generation sub-network 522 may be randomly initialized or obtained from a set word dictionary. The number of the contextual vectors is not limited, and may be 4, 8, 16, etc.
- [00149] According to the visual language model provided by an embodiment of the application, the prompt information generation network may be used to dynamically generate the category prompt information of each image category in the image category set based on the image feature of each image sample output by the image encoder network in the visual language model.
- [00150] Based on the visual language model, an embodiment of the application provides a model training method. FIG. 5B illustrates an implementation flowchart of a model training method provided by an embodiment of the application. As shown in FIG. 5B, the method includes S501 to S503 below.
- [00151] At S501, an image sample is processed by using a visual language model to be trained to obtain a classification result of the image sample.
- [00152] At S502, the cross entropy loss between the classification result and the category label of the image sample is determined.
- [00153] At S503, based on the cross entropy loss, at least one updating is performed on a network parameter of a mark extraction sub-network in the visual language model and at least one contextual vector adopted in an information generation sub-network to obtain

the trained visual language model.

- [00154] In the model training process, since the category prompt information of each image category is dynamically determined based on the image feature of the image sample, the text feature of each image category obtained based on the dynamically determined category prompt information can better reflect the feature of the image sample, so as to enhance the image recognition accuracy of the trained model. Moreover, compared with the static category prompt information, the dynamically determined category prompt information can also reduce the over-fitting of the image categories learned in the model training process. Thus, based on the dynamically determined category prompt information, the ability of recognizing image categories which are not involved in the model training process can be improved, and then the generalization ability of the trained visual language model in an image classification and recognition task can be improved. In some implementation modes, the trained model may be migrated to other cross domain image recognition tasks with the image sample used for training, and the performance of the migrated model in other cross domain image recognition tasks can be improved. Moreover, the network parameter of the mark extraction sub-network in the visual language model and the contextual vector may be learned, and the number of parameters to be learned in the model to be trained may be kept small, which can reduce the number of samples used in model training and enhance the training efficiency.
- [00155] The image processing method and the model training method provided by the embodiments of the application may be applied to any appropriate image recognition scene, such as a scene in which the category (such as human, cat, dog, rabbit and tiger) of an object contained in an image is recognized, a scene in which the category (such as indoor space, airport, station and shopping mall) of a scene represented by the image is recognized, a scene in which the state category (such as defective and indefective) of a target object in the image is recognized, a scene in which the posture category (such as stand, sit and lie) of a human body in the image is recognized, etc. The model training method and the image processing method provided by the embodiments of the application are described below by taking the scene in which the category of the object contained in the image is recognized as an example. The model training method may include the following operations.
- [00156] At S511, a training sample set including at least one image sample is obtained, where each image sample has a category label used to represent the category of an object contained in the image sample.
- [00157] At S512, classified processing is performed on the image feature by using a visual language model to be trained to obtain a classification result of the object contained in the image sample, and the cross entropy loss between the classification result and the category label of the image sample is determined.
- [00158] At S513, based on the cross entropy loss, at least one updating is performed on network parameter of a mark extraction sub-network in the visual language model and at least one contextual vector adopted in an information generation sub-network to obtain the trained visual language model.
- [00159] Any image including an object (namely an image to be processed) may be recognized by using the trained visual language model to obtain the category of the

object contained in the image. In the process of recognizing the image by using the trained visual language model, feature extraction may be performed on the image by using the trained image feature extraction network at first to obtain the image feature of the image; secondly, the image feature is processed by using the trained mark extraction sub-network to obtain the image mark information, and the category prompt information of each image category in the image category set is determined based on the image mark information and the trained contextual semantic information; the category prompt information of each image category in the image category set is determined based on the image feature; then, feature extraction is performed on each category of prompt information by using the trained text feature extraction network to obtain the text feature of each image category; finally, based on the text feature of each image category, classified processing is performed on the image feature to obtain the classification result of the image, namely the category of the object contained in the image. Here, each image category in the image category set may be an object category, and the object category in the image category set may be the object category involved in the image sample in the training sample set or the object category which is not involved in the image sample in the training sample set. In this way, in the process of recognizing the category of the object contained in the image, since the category prompt information of each image category is dynamically determined based on the image feature of the image, the text feature of each image category obtained based on the dynamically determined category prompt information can better reflect the feature of the image, so as to enhance the accuracy of recognizing the category of the object contained in the image, and the recognition ability of the object category which is not involved in the model training process can also be improved based on the dynamically determined category prompt information.

- [00160] FIG. 6 illustrates a composition structure diagram of an image processing apparatus provided by an embodiment of the application. As shown in FIG. 6, the image processing apparatus 600 includes an obtaining module 610, a first determination module 620, a second determination module 630 and a classification module 640.
- [00161] The obtaining module 610 is configured to obtain an image feature of an image to be processed.
- [00162] The first determination module 620 is configured to determine category prompt information of each image category in an image category set based on the image feature.
- [00163] The second determination module 630 is configured to determine a text feature of each image category based on each piece of category prompt information.
- [00164] The classification module 640 is configured to perform classified processing on the image feature based on the text feature of each image category to obtain a classification result of the image to be processed.
- [00165] In some embodiments, the first determination module is further configured to determine image mark information based on the image features, and determine the category prompt information of each image category in the image category set based on the image mark information and set contextual semantic information.
- [00166] In some embodiments, the first determination module is further configured to, for each image category in the image category set, determine the category prompt

information of the image category based on the image mark information, the contextual semantic information and a word embedding vector corresponding to the image category.

- [00167] In some embodiments, the contextual semantic information includes at least one contextual vector, the image mark information includes an image mark vector, and the category prompt information includes a category prompt vector. The first determination module is further configured to, for each contextual vector, determine an image contextual vector based on the image mark vector and the contextual vector, and determine the category prompt vector of the image category based on each image contextual vector and the word embedding vector.
- [00168] In some embodiments, the contextual semantic information includes at least one of: at least one randomly generated contextual vector; or at least one contextual vector obtained from a set word dictionary.
- [00169] In some embodiments, the obtaining module is further configured to perform feature extraction on the image to be processed by using the image feature extraction network in the model to obtain the image feature of the image to be processed. The first determination module is further configured to determine category prompt information of each image category in the image category set based on the image feature by using the prompt information generation network in the model. The second determination module is further configured to perform feature extraction on each piece of category prompt information by using the text feature extraction network in the model to obtain the text feature of each image category.
- [00170] In some embodiments, the prompt information generation network includes a mark extraction sub-network and an information generation sub-network. The first determination module is further configured to process the image feature by using the mark extraction sub-network to obtain image mark information, and determine the category prompt information of each image category in the image category set based on the image mark information and set contextual semantic information by using the information generation sub-network.
- [00171] In some embodiments, the classification result includes the probability that the image to be processed belongs to each image category. The classification module is further configured to determine the similarity between the image feature and the text feature of each image category, and determine the probability that the image to be processed belongs to each image category based on each similarity.
- [00172] FIG. 7 illustrates a composition structure diagram of a model training apparatus provided by an embodiment of the application. As shown in FIG. 7, the model training apparatus 700 includes a processing module 710 and an updating module 720.
- [00173] The processing module 710 is configured to process an image sample by utilizing a model to be trained to obtain a classification result of the image sample. Processing the image sample by utilizing the model may include: an image feature of the image sample is obtained; category prompt information of each image category in an image category set is determined based on the image feature; a text feature of each image category is determined based on each piece of category prompt information; and classified processing is performed on the image feature based on the text feature of each image category to obtain a classification result of the image sample.

- [00174] The updating module 720 is configured to perform at least one updating on network parameter of the model based on the classification result and a category label of the image sample to obtain the trained model.
- [00175] In some embodiments, the model includes an image feature extraction network, a prompt information generation network and a text feature extraction network. The processing module includes: a first extraction module configured to perform feature extraction on an image sample by using the image feature extraction network to obtain the image feature of the image sample; a third determination module configured to determine category prompt information of each image category in an image category set based on the image feature by using the prompt information generation network; a second extraction module configured to perform feature extraction on each piece of category prompt information by using the text feature extraction network to obtain a text feature of each image category; and a classification module configured to perform classified processing on the image feature based on the text feature of each image category to obtain the classification result of the image sample.
- [00176] In some embodiments, the updating module is further configured to perform at least one updating on a network parameter of the prompt information generation network based on the classification result and the category label of the image sample to obtain the trained model.
- [00177] In some embodiments, the prompt information generation network includes a mark extraction sub-network and an information generation sub-network. The third determination module is further configured to process the image feature by using the mark extraction sub-network to obtain image mark information, and determine category prompt information of each image category in the image category set based on the image mark information and contextual semantic information by using the information generation sub-network.
- [00178] In some embodiments, the updating module is further configured to perform at least one updating on a network parameter of the mark extraction sub-network and the contextual semantic information based on the classification result and the category label to obtain the trained model.
- [00179] In some embodiments, the updating module is further configured to determine a loss value based on the classification result and the category label of the image sample, and perform at least one updating on the network parameter of the model based on the loss value to obtain the trained model.
- [00180] In some embodiments, the apparatus further includes at least one of: a first initialization module configured to initialize the contextual semantic information based on at least one randomly generated contextual vector; and a second initialization module configured to initialize the contextual semantic information based on at least one contextual vector in a set word dictionary.
- [00181] The above descriptions about the apparatus embodiments are similar to descriptions about the method embodiments and beneficial effects similar to those of the method embodiments are achieved. Technical details undisclosed in the apparatus embodiments of the application may be understood with reference to the descriptions about the method embodiments of the application.

- [00182] It is to be noted that, in the embodiments of the application, when being implemented in form of software function module and sold or used as an independent product, the imaging processing method or the model training method may also be stored in a computer-readable storage medium. Based on such an understanding, the technical solutions of the embodiments of the application substantially or parts making contributions to the related art may be embodied in form of software product, and the software product is stored in a storage medium, including a plurality of instructions configured to enable a computer device (which may be a personal computer, a server, a network device and the like) to execute all or part of the method in each embodiment of the application. The storage medium includes: various media capable of storing program codes such as a flash disk, a mobile hard disk, a Read Only Memory (ROM), a magnetic disk or an optical disk. Therefore, the application is not limited to any specific combination of hardware and software.
- [00183] An embodiment of the application provides a computer device, which includes a memory and a processor. The memory stores a computer program capable of running on the processor. The processor implements the steps in the above method when executing the program.
- [00184] An embodiment of the application provides a computer readable storage medium, in which a computer program is stored. When executed by the processor, the computer program implements the steps in the above method.
- [00185] An embodiment of the application provides a computer program product, which may include a non-instantaneous computer-readable storage medium, in which a computer program is stored. When being read and executed by a computer, the computer program implements part or all steps in the above method. The computer program product may be specifically realized by means of hardware, software or a combination thereof. In an optional embodiment, the computer program product is specifically embodied as a computer storage medium, and in another optional embodiment, the computer program product is specifically embodied as software products, such as a Software Development Kit (SDK).
- [00186] It is to be pointed out here that the above descriptions about the storage medium, computer program product and device embodiments are similar to descriptions about the method embodiments and beneficial effects similar to those of the method embodiments are achieved. Technical details undisclosed in the storage medium, computer program product and device embodiments of the application may be understood with reference to the descriptions about the method embodiments of the application.
- [00187] It is to be noted that FIG. 8 illustrates a schematic diagram of a hardware entity of a computer device provided by an embodiment of the application. As shown in FIG. 8, the hardware entity of the computer device 800 includes a processor 801, a communication interface 802 and a memory 803.
- [00188] The processor 801 usually controls the overall operation of the computer device 800.
- [00189] The communication interface 802 may enable the computer device to communicate with other terminals or servers through the network.

- [00190] The memory 803 is configured to store an instruction and application executable for the processor 801, may also cache data (for example, image data, video data, voice communication data and video communication data) to be processed or having been processed by the processor 801 and each module in the computer device 800 and may be implemented through a flash or a Random Access Memory (RAM).
- [00191] It is to be understood that "one embodiment" and "an embodiment" mentioned in the whole specification mean that specific features, structures or characteristics related to the embodiment is included in at least one embodiment of the application. Therefore, "in one embodiment" or "in an embodiment" appearing at any place of the whole specification does not always refer to the same embodiment. In addition, these specific features, structures or characteristics may be combined in one or more embodiments in any proper manner. It should be understood that, in various embodiments of the present invention, a magnitude of a sequence number of each process does not mean an execution sequence and the execution sequence of each process should be determined by its function and an internal logic and should not form any limit to an implementation process of the embodiments of the application. The sequence numbers of the embodiments of the application are just for describing, instead of representing superiority-inferiority of the embodiments.
- [00192] It is to be noted that terms "include" and "contain" or any other variant thereof is intended to cover nonexclusive inclusions herein, so that a process, method, object or device including a series of components not only includes those components but also includes other components which are not clearly listed or further includes components intrinsic to the process, the method, the object or the device. Under the condition of no more limitations, a component defined by the statement "including a/an....." does not exclude existence of the same other components in a process, method, object or device including the component.
- [00193] In several embodiments provided by the application, it should be understood that the disclosed device and method may be implemented in another manner. The device embodiment described above is only schematic, and for example, division of the units is only logic function division, and other division manners may be adopted during practical implementation. For example, multiple units or components may be combined or integrated into another system, or some characteristics may be neglected or not executed. In addition, coupling or direct coupling or communication connection between each displayed or discussed component may be indirect coupling or communication connection, implemented through some interfaces, of the device or the units, and may be electrical and mechanical or adopt other forms.
- [00194] The units described as separate parts may or may not be physically separated, and parts displayed as units may or may not be physical units, and namely may be located in the same place, or may also be distributed to multiple network units. Part of all of the units may be selected according to a practical requirement to achieve the purposes of the solutions of the embodiments.
- [00195] In addition, each functional unit in each embodiment of the application may be integrated into a processing unit, each unit may also serve as an independent unit and two or more than two units may also be integrated into a unit. The integrated unit may be implemented in a hardware form and may also be implemented in form of hardware and software functional unit.

- [00196] Those of ordinary skill in the art should know that all or part of the steps of the method embodiment may be implemented by related hardware instructed through a program, the program may be stored in a computer-readable storage medium, and the program is executed to execute the steps of the method embodiment. The storage medium which may be a volatile storage medium includes: various media capable of storing program codes such as a mobile storage device, a ROM, a magnetic disk or a compact disc.
- [00197] Or, when being implemented in form of software function module and sold or used as an independent product, the integrated unit of the application may also be stored in a computer-readable storage medium. Based on such an understanding, the technical solutions of the application substantially or parts making contributions to the related art may be embodied in form of software product, and the computer software product is stored in a storage medium, including a plurality of instructions configured to enable a computer device (which may be a personal computer, a server, a network device and the like) to execute all or part of the method in each embodiment of the application. The storage medium includes: various media capable of storing program codes such as a mobile hard disk, a ROM, a magnetic disk or a compact disc.
- [00198] The above is only the specific implementation manner of the application and not intended to limit the scope of protection of the application. Any variations or replacements apparent to those skilled in the art within the technical scope disclosed by the application shall fall within the scope of protection of the application.

CLAIMS

1. An image processing method, comprising:
 - obtaining an image feature of an image to be processed;
 - determining category prompt information of each image category in an image category set based on the image feature;
 - determining a text feature of each the image category based on each the category prompt information; and
 - performing classified processing on the image feature based on the text feature of each the image category to obtain a classification result of the image to be processed.
2. The method according to claim 1, wherein determining the category prompt information of each the image category in the image category set based on the image features comprises:
 - determining image mark information based on the image feature; and
 - determining the category prompt information of each the image category in the image category set based on the image mark information and set contextual semantic information.
3. The method according to claim 2, wherein determining the category prompt information of each the image category in the image category set based on the image mark information and the set contextual semantic information comprises:
 - for each image category in the image category set, determining the category prompt information of the image category based on the image mark information, the contextual semantic information, and a word embedding vector corresponding to the image category.
4. The method according to claim 3, wherein the contextual semantic information comprises at least one contextual vector; the image mark information comprises an image mark vector; the category prompt information comprises a category prompt vector;
wherein determining the category prompt information of the image category based on the image mark information, the contextual semantic information, and the word embedding vector corresponding to the image category comprises:
 - for each the contextual vector, determining an image contextual vector based on the image mark vector and the contextual vector; and
 - determining the category prompt vector of the image category based on each the image contextual vector and the word embedding vector.
5. The method according to any one of claims 2 to 4, wherein the contextual semantic information comprises at least one of:
 - at least one randomly generated contextual vector; or

at least one contextual vector obtained from a set word dictionary.

6. The method according to any one of claims 1 to 5, wherein obtaining the image feature of the image to be processed comprises: performing feature extraction on the image to be processed by using an image feature extraction network in a model to obtain the image feature of the image to be processed;

wherein determining the category prompt information of each the image category in the image category set based on the image feature comprises: determining the category prompt information of each the image category in the image category set based on the image feature by using a prompt information generation network in the model;

wherein determining the text feature of each the image category based on each the category prompt information comprises: performing feature extraction on each the category prompt information by using a text feature extraction network in the model to obtain the text feature of each the image category.

7. The method according to claim 6, wherein the prompt information generation network comprises a mark extraction sub-network and an information generation sub-network;

wherein determining the category prompt information of each the image category in the image category set based on the image feature by using the prompt information generation network in the model comprises:

processing the image feature by using the mark extraction sub-network to obtain image mark information; and

determining the category prompt information of each the image category in the image category set based on the image mark information and the set contextual semantic information by using the information generation sub-network.

8. The method according to any one of claims 1 to 7, wherein the classification result comprises the probability that the image to be processed belongs to each the image category;

wherein performing classified processing on the image feature based on the text feature of each the image category to obtain the classification result of the image to be processed comprises:

determining the similarity between the image feature and the text feature of each the image category; and

determining the probability that the image to be processed belongs to each the image category based on each the similarity.

9. A model training method, comprising:

processing an image sample by using a model to be trained to obtain a classification result of the image sample, wherein processing the image sample by using the model comprises: obtaining an image feature of the image sample; determining category prompt

information of each image category in an image category set based on the image feature; determining a text feature of each the image category based on each the category prompt information; performing classified processing on the image feature based on the text feature of each the image category to obtain a classification result of the image sample; and

performing at least one updating on a network parameter of the model based on the classification result and a category label of the image sample to obtain the trained model.

10. The method according to claim 9, wherein the model comprises an image feature extraction network, a prompt information generation network and a text feature extraction network;

wherein processing the image sample by using the model to be trained to obtain the classification result of the image sample comprises:

performing feature extraction on the image sample by using the image feature extraction network to obtain the image feature of the image sample;

determining the category prompt information of each the image category in the image category set based on the image feature by using the prompt information generation network;

performing feature extraction on each the category prompt information by using the text feature extraction network to obtain the text feature of each the image category; and

performing classified processing on the image feature based on the text feature of each the image category to obtain the classification result of the image sample.

11. The method according to claim 10, wherein performing at least one updating on the network parameter of the model based on the classification result and the category label of the image sample to obtain the trained model comprises:

performing at least one updating on a network parameter of the prompt information generation network based on the classification result and the category label of the image sample to obtain the trained model.

12. The method according to claim 10 or 11, wherein the prompt information generation network comprises a mark extraction sub-network and an information generation sub-network;

wherein determining the category prompt information of each the image category in the image category set based on the image feature by using the prompt information generation network comprises:

processing the image feature by using the mark extraction sub-network to obtain image mark information; and

determining the category prompt information of each the image category in the image category set based on the image mark information and contextual semantic information by using the information generation sub-network.

13. The method according to claim 12, wherein performing at least one updating on the network parameter of the model based on the classification result and the category label of the image sample to obtain the trained model comprises:

performing at least one updating on a network parameter of the mark extraction sub-network and the contextual semantic information based on the classification result and the category label to obtain the trained model.

14. The method according to any one of claims 9 to 13, wherein performing at least one updating on the network parameter of the model based on the classification result and the category label of the image sample to obtain the trained model comprises:

determining a loss value based on the classification result and the category label of the image sample; and

performing at least one updating on the network parameter of the model based on the loss value to obtain the trained model.

15. The method according to any one of claims 9 to 14, further comprising at least one of:

initializing the contextual semantic information based on at least one randomly generated contextual vector; and

initializing the contextual semantic information based on at least one contextual vector obtained from a set word dictionary.

16. An image processing apparatus, comprising:

an obtaining module, configured to obtain an image feature of an image to be processed;

a first determination module, configured to determine category prompt information of each image category in an image category set based on the image feature;

a second determination module, configured to determine a text feature of each the image category based on each the category prompt information; and

a classification module, configured to perform classified processing on the image feature based on the text feature of each the image category to obtain a classification result of the image to be processed.

17. A model training apparatus, comprising:

a processing module, configured to process an image sample by using a model to be trained to obtain a classification result of the image sample, wherein processing the image sample by using the model comprises: obtaining a image feature of the image sample; determining category prompt information of each image category in an image category set based on the image feature; determining a text feature of each the image category based on each the category prompt information; and performing classified processing on the image

feature based on the text feature of each the image category to obtain a classification result of the image sample; and

an updating module, configured to perform at least one updating on a network parameter of the model based on the classification result and a category label of the image sample to obtain the trained model.

18. A computer device, comprising a memory and a processor, wherein the memory stores a computer program capable of running on the processor, wherein when executing the program, the processor implements the steps in the method of any one of claims 1 to 15.

19. A nonvolatile computer-readable storage medium, having a computer program stored thereon, wherein when executed by a processor, the computer program implements the steps in the method of any one of claims 1 to 15.

20. A computer program product, comprising a non-transitory computer-readable storage medium having a computer program stored thereon, wherein when read and executed by a computer, the computer program implements the steps in the method of any one of claims 1 to 15.