Claims

1. A method for training a scene recognition model, the scene recognition model comprising a core feature extraction layer, a global information feature extraction layer connected to the core feature extraction layer, a local supervised learning (LCS) module of at least one level with an attention mechanism, and a fully-connected decision layer;

wherein the method comprises:

acquiring parameters of the core feature extraction layer and the global information feature extraction layer by training based on a first scene label of a sample image and a standard cross-entropy loss;

training a weight parameter of the LCS module of each level, based on a loss value acquired by performing a pixel-by-pixel calculation on a feature map output from the LCS module of each level and the first scene label of the sample image; and

acquiring a parameter of the fully-connected decision layer by training based on the first scene label of the sample image and the standard cross-entropy loss;

wherein training the weight parameter of the LCS module of each level based on the loss value acquired by performing the pixel-by-pixel calculation on the feature map output from the LCS module of each level and the first scene label of the sample image comprises:

acquiring an importance weight of each channel using an activation function by an attention mechanism of a channel dimension;

acquiring a summary heat map by performing a weighted summation on the feature maps of the channels based on the importance weight of each channel; and

calculating the loss value pixel-by-pixel based on the summary heat map and the first scene label.

2. The method according to claim 1, wherein

the scene recognition model further comprises a branch expansion structure, the branch expansion structure comprising a convolutional layer and a local object association relationship learning module; and

the method further comprises:

training a weight parameter of a convolutional layer of at least one level of the branch expansion structure, based on a loss value acquired by performing a pixel-by-pixel calculation on a feature map output from a convolutional layer of the branch expansion structure and a second scene label of the sample image; and

acquiring a parameter of the local object association relationship learning module by training based on a loss function with a scene confidence regularization term, wherein the first scene label and the second scene label have different granularities.

3. The method according to claim 1, wherein the core feature extraction layer comprises a first-class packet multi-receptive field residual convolution module and a second-class packet multi-receptive field residual convolution module; wherein

the first-class packet multi-receptive field residual convolution module comprises a first packet, a second packet, and a third packet, wherein the first packet, the second packet, and the third packet have different convolution sizes, each of the first packet, the second packet, and the third packet comprises a residual calculation bypass structure, and each of the packets outputs a feature map by a convolution operation and a residual calculation, the feature map output from each of the packets being spliced and channel shuffled in a channel dimension and output to a next module upon convolutional fusion; and

the second-class packet multi-receptive field residual convolution module comprises a fourth packet, a fifth packet, and a sixth packet, wherein the fourth packet, the fifth packet, and the sixth packet have different convolution sizes, and each of the fifth packet and the sixth packet comprises a 1x1 convolution bypass structure and a residual calculation bypass structure, a feature map output from each of the packets being spliced and channel shuffled in the channel dimension and output to a next module upon convolutional fusion.

4. The method according to claim 1, wherein acquiring the parameters of the core feature extraction layer and the global information feature extraction layer by training based on the first scene label of the sample image and the standard cross-entropy loss comprises:

up-sampling feature maps of different levels in the core feature extraction layer by an inverse convolution operation with different expansion factors;

aligning the number of channels using a bilinear interpolation algorithm in a channel dimension;

summing and merging the feature map of at least one level channel-by-channel;

convolving and fusing the merged feature map group;

acquiring a global information feature vector by channel-by-channel global average pooling;

splicing the global information feature vector and a fully-connected layer FC feature vector; and

acquiring the parameters of the core feature extraction layer and the global information feature extraction layer by training based on the standard cross-entropy loss.

5. The method according to claim 1, wherein training the weight parameter of the LCS module of each level based on the loss value acquired by performing the pixel-by-pixel calculation on the feature map output from the LCS module of each level and the first scene label of the sample image further comprises:

calculating the loss value pixel-by-pixel based on the summary heat map, the first scene label, an object scene association importance, and an area of an object; and

training the weight parameter of the LCS module of each level based on the loss value.

6. The method according to claim 2, wherein the branch expansion structure is constructed by using a depth-wise separable convolutional residual block DW, in a main path of the residual block, a DW convolutional layer is used as a middle layer, and a 1x1 convolutional layer is used before and after the DW convolutional layer.

7. The method according to claim 2, wherein the local object association relationship learning module comprises a deformable convolutional layer, a convolutional layer, and an average pooling layer; wherein

the deformable convolution layer acquires a convolution kernel offset value of a current pixel position, a real effective position of a convolution kernel parameter is acquired by adding a current position of the convolution kernel parameter to the offset value, a feature image pixel value of the real effective position is acquired, and a feature map is output after a convolution operation and an average pooling operation.

8. A scene recognition method for a scene recognition model acquired by training according to the method as defined in any one of claims 1 to 7, the method comprising:

acquiring an image to be recognized; and

inputting the image to be recognized into a pre-trained scene recognition model and determining scene information corresponding to the image to be recognized based on the scene recognition model.

9. The method according to claim 8, further comprising:

in response to a scenario where the determined scene information corresponding to the image to be recognized belongs to offensive scene information and a moderation result of a machine moderation is that the image to be recognized is an offensive image, determining the image to be recognized to be an offensive image.

10. An apparatus for training a scene recognition model training, comprising:

a first training unit, configured to acquire parameters of a core feature extraction layer and a global information feature extraction layer by training based on a first scene label of a sample image and a standard cross-entropy loss;

a second training unit, configured to train a weight parameter of an LCS module of each level based on a loss value acquired by performing a pixel-by-pixel calculation on a feature map output from the LCS module of each level and the first scene label of the sample image; and

a third training unit, configured to acquire a parameter of a fully-connected decision layer by training based on the first scene label of the sample image and the standard cross-entropy loss;\

wherein the second training unit is further configured to:

acquire an importance weight of each channel using an activation function by an attention mechanism of a channel dimension;

acquire a summary heat map by performing a weighted summation on the feature maps of the channels based on the importance weight of each channel; and

calculate the loss value pixel-by-pixel based on the summary heat map and the first scene label.

11. A scene recognition apparatus for a scene recognition model acquired by training according to the apparatus according to claim 10, the apparatus comprising:

an acquisition unit, configured to acquire an image to be recognized; and

a recognition unit, configured to input the image to be recognized into a pre-trained scene recognition model and determine scene information corresponding to the image to be recognized based on the scene recognition model.

12. An electronic device, comprising: a processor and a memory, the memory storing one or more computer programs therein, wherein the processor, when loading and running the one or more computer programs stored in the memory, is caused to perform the steps of the method for training the scene recognition model as defined in any one of claims 1 to 7, or the steps of the scene recognition method as defined in claim 8 or 9.

13. A computer-readable storage medium, storing one or more computer programs therein, wherein the one or more computer programs, when loaded and run by a processor, cause the processor to perform the steps of the method for training the scene recognition model as defined in any one of claims 1 to 7, or the steps of the scene recognition method as defined in claim 8 or 9.