Claims

1. A method for embedding watermark information, comprising:

obtaining weight information of a target neural network model;

obtaining target watermark information; and

using the target watermark information and the weight information of the target neural network model to train an embedded neural network model, and updating the weight information of the target neural network model according to target watermark embedded data provided by the embedded neural network model to obtain the target neural network model embedded with the target watermark information, wherein using the target watermark information and the weight information of the target neural network model to train the embedded neural network model comprises:

obtaining random noise information;

using the weight information of the target neural network model as an instance of a first watermark training set, and using the target watermark information as a label of the first watermark training set;

using weight information of a reference neural network model as an instance of a second watermark training set, and using the random noise information as a label of the second watermark training set, the reference neural network being a neural network model without the target watermark information embedded; and

using the first watermark training set and the second watermark training set as a new training set, and updating the weight information of the embedded neural network model according to a model loss function of the embedded neural network until the model loss function converges.

2. The method according to claim 1, wherein using the target watermark information and the weight information of the target neural network model to train the embedded neural network model comprises:

obtaining weight information of the reference neural network model, the reference neural network model being a pre-trained model without the watermark information embedded, and the reference neural network model and the target neural network having a same structure; and

training the embedded neural network model according to the weight information of the reference neural network model, the random noise information, the weight information of the target neural network model and the target watermark information, wherein the random noise information and the target watermark information having a same structure.

3. The method according to claim 1, wherein updating the weight information of the target neural network model according to the target watermark embedded data provided by the embedded neural network model to obtain the target neural network model embedded with the target watermark information comprises:

inputting first weight information of the target neural network model into the embedded neural network model;

obtaining first target watermark embedded data outputted by the embedded neural network model, the first target watermark embedded data including watermark information extracted from the first weight information of the target neural network model by the embedded neural network model;

terminating the training of the embedded neural network model if the first target watermark embedded data indicates that the target watermark information has been embedded in the first weight information, and determining that a target neural network

model corresponding to the first weight information as the target neural network model embedded with the target watermark information.

4. The method according to claim 1, wherein updating the weight information of the target neural network model according to the target watermark embedded data provided by the embedded neural network model to obtain the target neural network model embedded with the target watermark information comprises:

inputting first weight information of the target neural network model into the embedded neural network model;

obtaining first target watermark embedded data outputted by the embedded neural network model. the first target watermark embedded data including watermark information extracted from the first weight information of the target neural network model by the embedded neural network model;

training the embedded neural network model if the first target watermark embedded data indicates that the target watermark information is not completely embedded in the first weight information to obtain an embedded neural network model that completes a first training, and providing the first target watermark embedded data to the target neural network model;

updating the weight information of the target neural network model according to the first target watermark embedded data to obtain second weight information of the target neural network model;

repeating the inputting, the obtaining and the training using the second weight information as the first weight information until the target neural network model that is completely embedded with the target watermark information and the target watermark information of the embedded neural network model that is able to be extracted from the

weight information of the target neural network model are obtained.

5. The method according to claim 4, wherein training the embedded neural network model to obtain the embedded neural network model that completes the first training if the first target watermark embedded data indicates that the target watermark information is not completely embedded in the target neural network model, comprises:

using the first weight information of the target neural network model and target watermark information as a third watermark training set, updating the weight information of the embedded neural network model according to the third watermark training set to obtain the embedded neural network model that completes the first training.

6. The method according to claim 4, wherein training the embedded neural network model to obtain the target neural network model that completes the first training if the first target watermark embedded data indicates that the target watermark information is not completely embedded in the target neural network model, comprises:

obtaining first weight information of the reference neural network model, the reference neural network model and the target neural network model having a same structure, wherein the random noise information is not completely embedded in the first weight information of the reference neural network model; and

using the first weight information of the reference neural network model, the random noise information, the first weight information of the target neural network model, and the target watermark information as a fourth watermark training set, and updating the weight information of the embedded neural network model according to the fourth watermark training set to obtain the embedded neural network model that completes the first training.

7. The method according to claim 1, wherein updating the weight information of the target neural network model according to the target watermark embedded data provided by the embedded neural network model to obtain the target neural network model embedded with the target watermark information comprises:

obtaining a regular term for embedding the target watermark information according to the target watermark embedded data;

adding the regular term for embedding the target watermark information on a basis of a corresponding model loss function when the target neural network model is trained using a conventional training set; and

updating the weight information of the target neural network based on the model loss function and the regular term for embedding the target watermark information, and obtaining the target neural network model embedded with the target watermark information.

8. The method according to claim 1, wherein the embedded neural network model is used for obtaining the target watermark embedded data according to the weight information of the target neural network model, and the target watermark embedded data is used for representing an embedding degree of embedded watermark information in the target neural network model.

9. An apparatus comprising:

one or more processors; and

memory storing executable instructions that, when executed by the one or more processors, cause the one or more processors to perform operations comprising:

obtaining weight information of a target neural network model;

obtaining target watermark information; and

using the target watermark information and the weight information of the target neural network model to train an embedded neural network model, and updating the weight information of the target neural network model according to target watermark embedded data provided by the embedded neural network model to obtain a target neural network model embedded with the target watermark information, wherein using the target watermark information and the weight information of the target neural network model to train the embedded neural network model comprises:

obtaining random noise information;

using the weight information of the target neural network model as an instance of a first watermark training set, and using the target watermark information as a label of the first watermark training set;

using weight information of a reference neural network model as an instance of a second watermark training set, and using the random noise information as a label of the second watermark training set, the reference neural network being a neural network model without the target watermark information embedded; and

using the first watermark training set and the second watermark training set as a new training set, and updating the weight information of the embedded neural network model according to a model loss function of the embedded neural network model until the model loss function converges.

10. The apparatus according to claim 9, wherein using the target watermark information and the weight information of the target neural network model to train the embedded neural network model comprises:

obtaining the weight information of the reference neural network model, the reference neural network model being a pre-trained model without the watermark information embedded, and the reference neural network model and the target neural network model having a same structure; and

training the embedded neural network model according to the weight information of the reference neural network model, the random noise information, the weight information of the target neural network model and the target watermark information, wherein the random noise information and the target watermark information having a same structure.

11. The apparatus according to claim 9, wherein updating the weight information of the target neural network model according to the target watermark embedded data provided by the embedded neural network model to obtain the target neural network model embedded with the target watermark information comprises:

Inputting first weight information of the target neural network model into the embedded neural network model;

obtaining first target watermark embedded data outputted by the embedded neural network model, the first target watermark embedded data including watermark information extracted from the first weight information of the target neural network model by the embedded neural network model;

terminating the training of the embedded neural network model if the first target watermark embedded data indicates that the target weight information has been embedded in the first weight information, and determining that a target neural network model corresponding to the first weight information as the target neural network model embedded with the watermark information.

12. The apparatus according to claim 9, wherein updating the weight information of the target neural network model according to the target watermark embedded data provided by the embedded neural network model to obtain the target neural network model embedded with the target watermark information comprises:

obtaining a regular term for embedding the target watermark information according to the target watermark embedded data;

adding the regular term for embedding the target watermark information on a basis of a corresponding model loss function when the target neural network model is trained using a conventional training set; and

updating the weight information of the target neural network model based on the model loss function and the regular term for embedding the target watermark information, and obtaining the target neuron network model embedded with the target watermark information.

13. The apparatus according to claim 9, wherein the embedded neural network model is used for obtaining the target watermark embedded data according to the weight information of the target neural network model, and the target watermark embedded data is used for representing an embedding degree of embedded watermark information in the target neural network model.