

CLAIMS

What is claimed is:

1. A model-based prediction method performed by a machine learning system, the system including a plurality of machine learning models and a plurality of embedding models for converting an input tensor of a machine learning model, the embedding models being deployed separately from the machine learning models, each embedding model being deployed in a respective embedding running environment of a plurality of embedding running environments, wherein each embedding model is a copy of a particular embedding model, and each machine learning model being deployed in a respective model running environment of a plurality of model running environments, wherein each model running environment is configured to interact with an embedding running environment, the method comprising:

receiving, by a model running environment of the plurality of model running environments, an input tensor of the machine learning model;

sending, by the model running environment, a table query request to an embedding running environment of the plurality of embedding running environments, the table query request including the input tensor, to request low-dimensional conversion of the input tensor;

obtaining, by the embedding running environment, a plurality of vectors based on the table query request and the input tensor;

generating a table query result based on the obtained plurality of vectors, wherein the table query result is generated, at least in part, by combining the obtained plurality of vectors into a single vector;

receiving, by the model running environment, the table query result returned by the embedding running environment, the table query result being obtained by the embedding running environment by querying an embedding table for the low-dimensional conversion that is associated with the machine learning model based on the input tensor; and

inputting, by the model running environment, the table query result into the machine learning model, and executing the machine learning model to complete model-based prediction.

2. The method according to claim 1, wherein the model running environment is a physical execution unit or a virtual execution unit, and the embedding running environment is a physical execution unit or a virtual execution unit.

3. The method according to claim 1, wherein the machine learning system includes at least one embedding running environment and at least one model running environment, each embedding running environment implementing a single embedding model, and each model running environment implementing a single machine learning model.

4. A machine learning method performed by a machine learning system executing on a plurality of running environments including a plurality of model running environments and a plurality of embedding running environments, wherein each model running environment is configured to interact with an embedding running environment and wherein each embedding running environment includes a copy of a particular embedding model, the method comprising:

receiving, by a model running environment of the plurality of model running environments, an input to a machine learning model implemented on the model running environment;

sending, from the model running environment to an embedding running environment of the plurality of embedding environments, a request for converting the input;

receiving, by the model running environment, a result returned from the embedding running environment, the result including a low-dimensional representation of the input; and

feeding, by the model running environment, the low-dimensional representation into the machine learning model to perform model-based prediction.

5. The method according to claim 4, wherein the sending, from the model running environment to the embedding running environment, a request for converting the input includes:

sending a local request for converting the input, wherein the embedding running environment and the model running environment are located on a same physical node; or

sending a remote request for converting the input, wherein the embedding running environment and the model running environment are located on different physical nodes.

6. The method according to claim 4, wherein the plurality of running environments includes at least two model running environments and different hardware resources are configured for different model running environments, the hardware resources being adapted to running requirements of machine learning models in the model running environments.

7. The method according to claim 6, wherein the hardware resource includes at least one of a central processing unit or a hardware accelerator.

8. The method of claim 7, wherein the hardware accelerator includes at least one of a graphics processing unit, a field-programmable gate array, or an application-specific integrated circuit chip designed for a specific purpose.

9. The method according to claim 4, wherein the machine learning model includes at least one of a deep neural network model, a Wide & Deep model, or a Deep Factorization Machine model.

10. A machine learning system comprising a plurality of embedding running environments and a plurality of model running environments, , wherein each embedding running environment of the plurality of embedding running environments includes a copy of a particular embedding model and wherein each model running environment of the plurality of model running environments is configured to interact with an embedding running environment, and wherein a machine learning model is deployed in each model running environment of the plurality of model running environments;

wherein a model running environment of the plurality of model running environments is configured to receive an input for the machine learning model, send a table query request including the input to an embedding running environment of the plurality of embedding running environments, receive a response including a low-dimensional converted value of the input from the embedding running environment, and feed the low-dimensional converted value into the machine learning model to execute model-based prediction; and

wherein the embedding running environment is configured to perform embedding query based on the input to obtain the low-dimensional converted value by obtaining a plurality of vectors based on the input and generating the response based, at least in part, on combining the plurality of vectors into a single vector, and send the response including the low-dimensional converted value back to the model running environment.

11. The system according to claim 10, wherein the embedding running environment and the model running environment each are a physical execution unit or a virtual execution unit.

12. The system according to claim 10, comprising at least one embedding running environment and at least one model running environment, each embedding running environment implementing at least one embedding model, and each model running environment implementing at least one machine learning model.

13. The system according to claim 12, wherein different hardware resources are configured for different model running environments, the hardware resources being adapted to running requirements of machine learning models in the model running environments.

14. A non-transitory storage medium storing contents that, when executed by one or more processors, cause the one or more processors to perform actions comprising:

receiving, by a model running environment of a plurality of model running environments included in a machine learning system, an input to a machine learning model implemented on the model running environment, wherein each model running environment is configured to interact with an embedding running environment;

sending, from the model running environment to an embedding running environment of a plurality of embedding running environments included in the machine learning system, a request for converting the input, wherein each embedding running environment of the plurality of embedding running environments includes a copy of a particular embedding model;

obtaining, by the embedding running environment, a plurality of vectors based on the table query request and the input;

generating a table query result based on the obtained plurality of vectors, wherein the table query result is generated, at least in part, by combining the obtained plurality of vectors into a single vector;

receiving, by the model running environment, a result returned from the embedding running environment, the result including a low-dimensional representation of the input; and

feeding, by the model running environment, the low-dimensional representation into the machine learning model to perform model-based prediction.

15. The storage medium according to claim 14, wherein the model running environment is a physical execution unit or a virtual execution unit, and the embedding running environment is a physical execution unit or a virtual execution unit.

16. The storage medium according to claim 14, wherein the sending, from the model running environment to the embedding running environment, a request for converting the input includes:

sending a local request for converting the input, wherein the embedding running environment and the model running environment are located on a same physical node; or

sending a remote request for converting the input, wherein the embedding running environment and the model running environment are located on different physical nodes.

17. The storage medium according to claim 14, wherein the machine learning system includes a plurality of model running environments and different hardware resources are configured for different model running environments, the hardware resources being adapted to running requirements of machine learning models in the model running environments.

18. The storage medium according to claim 17, wherein the hardware resource includes at least one of a central processing unit or a hardware accelerator.