A METHOD FOR AI BASED LOAD PREDICTION

TECHNICAL FIELD

This disclosure is directed to exchanging Artificial Intelligence (AI) computing information for load prediction within a wireless communication network.

BACKGROUND

Self-optimization Network (SON) has become a key feature in mobile communication systems. In particular, automatic Mobility Load Balancing (MLB) is an important function in SON. For example, with MLB in cellular networks, the communication system may evenly distribute cell traffic load and offload mobile users from one cell to another. Such optimization aims at increasing the system capacity and reliability as well as improving user experience. In addition, this function is able to minimize human intervention in the network management and optimization tasks.

SUMMARY

This disclosure describes methods and systems for exchanging AI computing information between various network elements in a wireless communication systems.

In ~~one example implementation~~a first aspect of the present invention, a method performed by a network element of a wireless communication system for exchanging AI computing information is disclosed. The method include sending, by a first network element of a wireless communication network, a first message for load prediction to a second network element of the wireless communication network, wherein the first message comprises at least one of an input to a machine learning model for load prediction of the first network element or model configuration information of the machine learning model, wherein the input to the machine learning model comprises at least one of: current load information of a current cell associated with and served by the first network element; current load information of a neighbor cell of the current cell; historical load information of the current cell; or historical load information of the neighbor cell.

In a second aspect of the present invention, there is disclosed a first network element comprising one or more processors and one or more memories, wherein the one or more processors are configured to read computer code from the one or more memories to: send a first message for load prediction to a second network element, wherein the first message comprises at least one of an input to a machine learning model for load prediction of the first network element or model configuration information of the machine learning model; receive a response message to the first message from the second network element; wherein the input to the machine learning model comprises at least one of: current load information of a current cell associated with and served by the first network element; current load information of a neighbor cell of the current cell; historical load information of the current cell; or historical load information of the neighbor cell; wherein the model configuration information of the machine learning model comprises at least one of:a machine learning model type; an AI algorithm information; information of a hardware platform on which the machine learning model is run; or training configuration information of the machine learning model, wherein the first message comprises one of: a Next-Generation Radio Access Network (NG-RAN) Node Configuration Update message; a Next generation nodeB Distribution Unit (gNB-DU) Configuration Update message; a Next generation nodeB Central Unit (gNB-CU) Configuration Update message; an EN-DC Configuration Update message; or an AI Configuration Update message; and wherein the response message comprises one of: a NG-RAN Node Configuration Update Acknowledge message; an AI Configuration Update Acknowledge message; a gNB-DU Configuration Update Acknowledge message; a gNB-CU Configuration Update Acknowledge message; or an EN-DC Configuration Update Acknowledge message.

In a third aspect of the present invention, there is disclosed a non-transitory storage medium for storing computer readable instructions, the computer readable instructions, when executed by a processor in a first network element, causing the processor to: send a first message for load prediction to a second network element, wherein the first message comprises at least one of an input to a machine learning model for load prediction of the first network element or model configuration information of the machine learning model; receive a response message to the first message from the second network element; wherein the input to the machine learning model comprises at least one of: current load information of a current cell associated with and served by the first network element; current load information of a neighbor cell of the current cell; historical load information of the current cell; or historical load information of the neighbor cell; wherein the model configuration information of the machine learning model comprises at least one of: a machine learning model type; an AI algorithm information; information of a hardware platform on which the

machine learning model is run; or training configuration information of the machine learning model, wherein the first message comprises one of: a Next-Generation Radio Access Network (NG-RAN) Node Configuration Update message; a Next generation nodeB Distribution Unit (gNB-DU) Configuration Update message; a Next generation nodeB Central Unit (gNB-CU) Configuration Update message; an EN-DC Configuration Update message; or an AI Configuration Update message; and wherein the response message comprises one of: a NG-RAN Node Configuration Update Acknowledge message; an AI Configuration Update Acknowledge message; a gNB-DU Configuration Update Acknowledge message; a gNB-CU Configuration Update Acknowledge message; or an EN-DC Configuration Update Acknowledge message.

In the implementation above, the input to the machine learning model includes at least one of: current load information of a current cell associated with and served by the first network element; current load information of a neighbor cell of the current cell; historical