SUPPORTING FEDERATED LEARNING IN A COMMUNICATION SYSTEM

TECHNICAL FIELD

[0001] The present disclosure relates to wireless communications, and more specifically to network analytics in wireless communication networks.

BACKGROUND

[0002] A wireless communications system may include one or multiple network communication devices, such as base stations, which may support wireless communications for one or multiple user communication devices, which may be otherwise known as user equipment (UE), or other suitable terminology. The wireless communications system may support wireless communications with one or multiple user communication devices by utilizing resources of the wireless communication system (e.g., time resources (e.g., symbols, slots, subframes, frames, or the like) or frequency resources (e.g., subcarriers, carriers, or the like)). Additionally, the wireless communications system may support wireless communications across various radio access technologies including third generation (3G) radio access technology, fourth generation (4G) radio access technology, fifth generation (5G) radio access technology, among other suitable radio access technologies beyond 5G (e.g., sixth generation (6G)).

SUMMARY

[0003] An article "a" before an element is unrestricted and understood to refer to "at least one" of those elements or "one or more" of those elements. The terms "a," "at least one," "one or more," and "at least one of one or more" may be interchangeable. As used herein, including in the claims, "or" as used in a list of items (e.g., a list of items prefaced by a phrase such as "at least one of" or "one or more of" or "one or both of") indicates an inclusive list such that, for example, a list of at least one of A, B, or C means A or B or C or AB or AC or BC or ABC (i.e., A and B and C). Also, as used herein, the phrase "based on" shall not be construed as a reference to a closed set of conditions. For example, an example step that is described as "based on condition A" may be based on both a condition A and a condition B without departing from the scope of the present disclosure. In other words, as

used herein, the phrase “based on” shall be construed in the same manner as the phrase “based at least in part on”. Further, as used herein, including in the claims, a “set” may include one or more elements.

[0004] Some implementations of the method and apparatuses described herein may further include a network entity for wireless communication supporting network data analytics, the network entity comprising instructions executable by at least one processor to cause the network entity to: obtain a request message to provide an inference output for a network analytics operation identified by an analytic identifier; output an inference request message to each of a set of one or more model training functions, set model training functions being participants in a federated learning training process for a trained model corresponding to the network analytics operation; obtain an inference output from each function of said set; output a response message comprising an aggregate inference output, said aggregate inference output being based on a respective contribution weight, in the federated learning training process, for each said function.

[0005] The instructions may cause the network entity to: obtain a model information message from a second network entity, the model information message comprising: an identifier of the model; identifiers of one or more functions to participate in the federated learning training process; and a coordination identifier associated with a current instance of the federated learning training process.

[0006] The model information message may comprise information identifying each of the model training functions as having one or more roles in the federated training process, said one or more roles being one or more of a set comprising a federated learning server, a federated learning active participant, and a federated learning passive participant.

[0007] One or more of said model training functions may comprise a network data analytics function (NWDAF). Alternatively or additionally, one or more of said model training functions may comprise an application function (AF).

[0008] The instructions may cause the network entity to obtain the model information message based on the analytics identifier.

[0009] The obtaining the model information message may comprise outputting, to the second network entity, a model information request message comprising the analytics identifier.

[0010] The second network entity may comprise at least one of a network repository function and an analytics data repository function.

[0011] The network entity may comprise at least one of: a network data analytics function (NWDAF), said NWDAF supporting an analytics logical function; and a NWDAF configured to control federated learning processes for the analytic identifier.

[0012] The instructions may cause the network entity to receive the request message from a consumer entity within the communications network.

[0013] The consumer entity may be a consumer analytics logical function.

[0014] The instructions may cause the network entity to output said response message to the consumer entity.

[0015] Said inference output may be said aggregate inference output, and the instructions may cause the network entity to obtain said aggregate inference output from a third network entity.

[0016] The third network entity may comprise a Model Training Logical Function (MTLF), said MTLF being a controller of the federated learning process for the said model.

[0017] Alternatively or additionally, the third network entity may comprise an AF.

[0018] Said inference output may comprise respective inference outputs from each of said one or more functions. The instructions may cause the network entity to: obtain each said inference output from the respective function of said one or more functions; and determine the aggregate inference output based on each said inference output.

[0019] The instructions may cause the network entity to: obtain said respective contribution weights; and determine the aggregate inference output based additionally on said respective contribution weights.

[0020] Said one or more network analytics operations may comprise at least one statistical analysis operation to be performed on the network.

[0021] Said one or more network analytics operations may comprise at least one prediction operation to be performed in respect of behavior of one or more entities within said communication network.

[0022] Said prediction operation may comprise at least one of: a prediction of a parameter associated with at least one entity within the communication network, said parameter optionally being a load of the at least one entity, said at least one entity optionally comprising at least one of a network entity and a UE of the communication network; and a prediction of a movement characteristic of a UE in the communication network, said movement characteristic optionally comprising a location of the UE.

[0023] Each of said one or more functions may be a network function for providing analytics data associated with one or more UEs within the communication network.

[0024] In some implementations of the methods and apparatuses described herein, a method is performed by a network entity of a wireless communication network, the method comprising: obtaining a request message to provide an inference output for a network analytics operation identified by an analytic identifier; outputting an inference request message to each of a set of one or more model training functions, set model training functions being participants in a federated learning training process for a trained model corresponding to the network analytics operation; obtaining an inference output from each function of said set; outputting a response message comprising an aggregate inference output, said aggregate inference output being based on a respective contribution weight, in the federated learning training process, for each said function.

[0025] In some implementations of the methods and apparatuses described herein, a communication network system comprises: a first network entity for wireless communication in the communication network; and a second network entity for wireless communication in the communication network, wherein: the first network entity is configured to: obtain a first request message to provide an inference output for a network analytics operation identified by an analytic identifier; output a second request message to

the second network entity to perform an inference operation in respect of the network analytics operation, the second network entity being configured to: output a third request message to each of one or more functions to participate in a federated learning training process for a model to perform said network analytics operation; obtain an inference output from each of said one or more functions; output to the first network entity an aggregated inference output based on said obtained inference outputs and respective contribution weights, in the federated learning training process, for each said function, and the first network entity being configured to output a response message comprising said aggregate inference output.

[0026] The second network entity may be configured to provide training data (for example gradients and/or losses) based on said inference outputs to said one or more functions for training said model.

[0027] The first network entity may be a NWDAF supporting AnLF.

[0028] The second network entity may be a federated learning server.

BRIEF DESCRIPTION OF THE DRAWINGS

[0029] Figure 1 illustrates an example of a wireless communications system in accordance with aspects of the present disclosure.

[0030] Figures 2 and 3 depict a network architecture.

[0031] Figures 4A and 4B illustrate federated learning techniques.

[0032] Figures 5A and 5B depict communication flows for federated learning.

[0033] Figure 6 depicts architectures for federated learning, according to an example.

[0034] Figures 7A and 7B depict process flows for federated learning, according to examples.

[0035] Figure 8 illustrates an example of a network equipment (NE) 800 in accordance with aspects of the present disclosure.

[0036] Figure 9 illustrates a flowchart of a method performed by a NE in accordance with aspects of the present disclosure.

[0037] Figure 10 schematically depicts a communication network in accordance with aspects of the present disclosure.

DETAILED DESCRIPTION

[0038] Some wireless communication systems, including one or more network entities may support network analytics operations (e.g., process, action), for example for performing statistical analyses of activity in these wireless communication systems. For example, analysis may be performed on parameters such as signal quality, traffic volumes, and the like. In some other examples, the network analytics operation may include the one or more network entities performing a prediction (e.g., estimation, determination, forecast) of a characteristic of the wireless communication systems (e.g., networks). For example, the one or more network entities may perform a prediction of a location (e.g., position) of a user equipment (UE) at a given time, or a prediction of a load (e.g., data traffic volume) for the one or more network entities at a given time in the future.

[0039] The present disclosure provides for improvements to the performance, for example the accuracy and/or processing, of such network analytics. Various aspects of the present disclosure relate to using one or more learning techniques, for example, a vertical federated learning technique, to perform network analytics. Federated learning techniques, described in more detail below, are techniques in which multiple entities collaborate to train a model, whilst their data is not centralized. It should be understood that other learning techniques (e.g., machine learning model, artificial intelligence models, or the like) may be used for support and enabling the one or more network entities to support network analytics. Some aspects of the present disclosure relate to control of the federated learning techniques (e.g., process) by one or more network entities, in order to effectively provide improved network analytics. This control, which is described in more detail below, may include transmitting messages to participants in the process to coordinate their training of a model. Such use of learning techniques can provide improved statistical analysis, and improved prediction quality, thereby improving the accuracy of network analytics. Improved network performance can thus be achieved.

[0040] Aspects of the present disclosure are described in the context of a wireless communications system.

[0041] Figure 1 illustrates an example of a wireless communications system 100 in accordance with aspects of the present disclosure. The wireless communications system 100 may include one or more NE 102, one or more UE 104, and a core network (CN) 106. The wireless communications system 100 may support various radio access technologies. In some implementations, the wireless communications system 100 may be a 4G network, such as an LTE network or an LTE-Advanced (LTE-A) network. In some other implementations, the wireless communications system 100 may be a NR network, such as a 5G network, a 5G-Advanced (5G-A) network, or a 5G ultrawideband (5G-UWB) network. In other implementations, the wireless communications system 100 may be a combination of a 4G network and a 5G network, or other suitable radio access technology including Institute of Electrical and Electronics Engineers (IEEE) 802.11 (Wi-Fi), IEEE 802.16 (WiMAX), IEEE 802.20. The wireless communications system 100 may support radio access technologies beyond 5G, for example, 6G. Additionally, the wireless communications system 100 may support technologies, such as time division multiple access (TDMA), frequency division multiple access (FDMA), or code division multiple access (CDMA), etc.

[0042] The one or more NE 102 may be dispersed throughout a geographic region to form the wireless communications system 100. One or more of the NE 102 described herein may be or include or may be referred to as a network node, a base station, a network element, a network function, a network entity, a radio access network (RAN), a NodeB, an eNodeB (eNB), a next-generation NodeB (gNB), or other suitable terminology. An NE 102 and a UE 104 may communicate via a communication link, which may be a wireless or wired connection. For example, an NE 102 and a UE 104 may perform wireless communication (e.g., receive signaling, transmit signaling) over a Uu interface.

[0043] An NE 102 may provide a geographic coverage area for which the NE 102 may support services for one or more UEs 104 within the geographic coverage area. For example, an NE 102 and a UE 104 may support wireless communication of signals related to services (e.g., voice, video, packet data, messaging, broadcast, etc.) according to one or

multiple radio access technologies. In some implementations, an NE 102 may be moveable, for example, a satellite associated with a non-terrestrial network (NTN). In some implementations, different geographic coverage areas 112 associated with the same or different radio access technologies may overlap, but the different geographic coverage areas may be associated with different NE 102.

[0044] The one or more UE 104 may be dispersed throughout a geographic region of the wireless communications system 100. A UE 104 may include or may be referred to as a remote unit, a mobile device, a wireless device, a remote device, a subscriber device, a transmitter device, a receiver device, or some other suitable terminology. In some implementations, the UE 104 may be referred to as a unit, a station, a terminal, or a client, among other examples. Additionally, or alternatively, the UE 104 may be referred to as an Internet-of-Things (IoT) device, an Internet-of-Everything (IoE) device, or machine-type communication (MTC) device, among other examples.

[0045] A UE 104 may be able to support wireless communication directly with other UEs 104 over a communication link. For example, a UE 104 may support wireless communication directly with another UE 104 over a device-to-device (D2D) communication link. In some implementations, such as vehicle-to-vehicle (V2V) deployments, vehicle-to-everything (V2X) deployments, or cellular-V2X deployments, the communication link 114 may be referred to as a sidelink. For example, a UE 104 may support wireless communication directly with another UE 104 over a PC5 interface.

[0046] An NE 102 may support communications with the CN 106, or with another NE 102, or both. For example, an NE 102 may interface with other NE 102 or the CN 106 through one or more backhaul links (e.g., S1, N2, N3, or network interface). In some implementations, the NE 102 may communicate with each other directly. In some other implementations, the NE 102 may communicate with each other indirectly (e.g., via the CN 106). In some implementations, one or more NE 102 may include subcomponents, such as an access network entity, which may be an example of an access node controller (ANC). An ANC may communicate with the one or more UEs 104 through one or more other access network transmission entities, which may be referred to as a radio heads, smart radio heads, or transmission-reception points (TRPs).

[0047] The CN 106 may support user authentication, access authorization, tracking, connectivity, and other access, routing, or mobility functions. The CN 106 may be an evolved packet core (EPC), or a 5G core (5GC), which may include a control plane entity that manages access and mobility (e.g., a mobility management entity (MME), an access and mobility management functions (AMF)) and a user plane entity that routes packets or interconnects to external networks (e.g., a serving gateway (S-GW), a Packet Data Network (PDN) gateway (P-GW), or a user plane function (UPF)). In some implementations, the control plane entity may manage non-access stratum (NAS) functions, such as mobility, authentication, and bearer management (e.g., data bearers, signal bearers, etc.) for the one or more UEs 104 served by the one or more NE 102 associated with the CN 106.

[0048] The CN 106 may communicate with a packet data network over one or more backhaul links (e.g., via an S1, N2, N3, or another network interface). The packet data network may include an application server. In some implementations, one or more UEs 104 may communicate with the application server. A UE 104 may establish a session (e.g., a protocol data unit (PDU) session, or the like) with the CN 106 via an NE 102. The CN 106 may route traffic (e.g., control information, data, and the like) between the UE 104 and the application server using the established session (e.g., the established PDU session). The PDU session may be an example of a logical connection between the UE 104 and the CN 106 (e.g., one or more network functions of the CN 106).

[0049] In the wireless communications system 100, the NEs 102 and the UEs 104 may use resources of the wireless communications system 100 (e.g., time resources (e.g., symbols, slots, subframes, frames, or the like) or frequency resources (e.g., subcarriers, carriers)) to perform various operations (e.g., wireless communications). In some implementations, the NEs 102 and the UEs 104 may support different resource structures. For example, the NEs 102 and the UEs 104 may support different frame structures. In some implementations, such as in 4G, the NEs 102 and the UEs 104 may support a single frame structure. In some other implementations, such as in 5G and among other suitable radio access technologies, the NEs 102 and the UEs 104 may support various frame structures (i.e., multiple frame structures). The NEs 102 and the UEs 104 may support various frame structures based on one or more numerologies.

[0050] One or more numerologies may be supported in the wireless communications system 100, and a numerology may include a subcarrier spacing and a cyclic prefix. A first numerology (e.g., μ=0) may be associated with a first subcarrier spacing (e.g., 15 kHz) and a normal cyclic prefix. In some implementations, the first numerology (e.g., μ=0) associated with the first subcarrier spacing (e.g., 15 kHz) may utilize one slot per subframe. A second numerology (e.g., μ=1) may be associated with a second subcarrier spacing (e.g., 30 kHz) and a normal cyclic prefix. A third numerology (e.g., μ=2) may be associated with a third subcarrier spacing (e.g., 60 kHz) and a normal cyclic prefix or an extended cyclic prefix. A fourth numerology (e.g., μ=3) may be associated with a fourth subcarrier spacing (e.g., 120 kHz) and a normal cyclic prefix. A fifth numerology (e.g., μ=4) may be associated with a fifth subcarrier spacing (e.g., 240 kHz) and a normal cyclic prefix.

[0051] A time interval of a resource (e.g., a communication resource) may be organized according to frames (also referred to as radio frames). Each frame may have a duration, for example, a 10 millisecond (ms) duration. In some implementations, each frame may include multiple subframes. For example, each frame may include 10 subframes, and each subframe may have a duration, for example, a 1 ms duration. In some implementations, each frame may have the same duration. In some implementations, each subframe of a frame may have the same duration.

[0052] Additionally or alternatively, a time interval of a resource (e.g., a communication resource) may be organized according to slots. For example, a subframe may include a number (e.g., quantity) of slots. The number of slots in each subframe may also depend on the one or more numerologies supported in the wireless communications system 100. For instance, the first, second, third, fourth, and fifth numerologies (i.e., μ=0, μ=1, μ=2, μ=3, μ=4) associated with respective subcarrier spacings of 15 kHz, 30 kHz, 60 kHz, 120 kHz, and 240 kHz may utilize a single slot per subframe, two slots per subframe, four slots per subframe, eight slots per subframe, and 16 slots per subframe, respectively. Each slot may include a number (e.g., quantity) of symbols (e.g., OFDM symbols). In some implementations, the number (e.g., quantity) of slots for a subframe may depend on a numerology. For a normal cyclic prefix, a slot may include 14 symbols. For an extended cyclic prefix (e.g., applicable for 60 kHz subcarrier spacing), a slot may include 12

symbols. The relationship between the number of symbols per slot, the number of slots per subframe, and the number of slots per frame for a normal cyclic prefix and an extended cyclic prefix may depend on a numerology. It should be understood that reference to a first numerology (e.g., μ=0) associated with a first subcarrier spacing (e.g., 15 kHz) may be used interchangeably between subframes and slots.

[0053] In the wireless communications system 100, an electromagnetic (EM) spectrum may be split, based on frequency or wavelength, into various classes, frequency bands, frequency channels, etc. By way of example, the wireless communications system 100 may support one or multiple operating frequency bands, such as frequency range designations FR1 (410 MHz – 7.125 GHz), FR2 (24.25 GHz – 52.6 GHz), FR3 (7.125 GHz – 24.25 GHz), FR4 (52.6 GHz – 114.25 GHz), FR4a or FR4-1 (52.6 GHz – 71 GHz), and FR5 (114.25 GHz – 300 GHz). In some implementations, the NEs 102 and the UEs 104 may perform wireless communications over one or more of the operating frequency bands. In some implementations, FR1 may be used by the NEs 102 and the UEs 104, among other equipment or devices for cellular communications traffic (e.g., control information, data). In some implementations, FR2 may be used by the NEs 102 and the UEs 104, among other equipment or devices for short-range, high data rate capabilities.

[0054] FR1 may be associated with one or multiple numerologies (e.g., at least three numerologies). For example, FR1 may be associated with a first numerology (e.g., μ=0), which includes 15 kHz subcarrier spacing; a second numerology (e.g., μ=1), which includes 30 kHz subcarrier spacing; and a third numerology (e.g., μ=2), which includes 60 kHz subcarrier spacing. FR2 may be associated with one or multiple numerologies (e.g., at least 2 numerologies). For example, FR2 may be associated with a third numerology (e.g., μ=2), which includes 60 kHz subcarrier spacing; and a fourth numerology (e.g., μ=3), which includes 120 kHz subcarrier spacing.

[0055] In the wireless communication system 100, a NE 102 may be configured (as described in more detail below) to control a federated learning process to perform network analytics.

[0056] Figure 2 illustrates a network architecture that may allow data collection relating to the use of network data analytics functions (NWDAFs). In some cases, an analytics logical function (AnLF) requests a trained machine learning (ML) model from the model training logical function (MTLF) by including in the request an Analytic ID corresponding to the Analytic ID requested by an analytics consumer. Information on the available Analytic ID(s) supported by an AnLF is provided in 3GPP TS 23.288. The MTLF trains an ML model corresponding to the Analytic ID requested by collecting data from one or more data sources (NF, OAM or UEs).

[0057] The functionality of a NWDAF, such as that of Figure 1, may be divided between an AnLF and a MTLF. Figure 3 illustrates an example of such a division of functionality.

[0058] In Figure 3, MTLFs each train a ML model for performing a respective network analytic. Each MTLF may be associated with an analytic ID. Each such analytic ID identifies a particular network analytic. A given ML model may be configured by a ML model designer.

[0059] Each MTLF may trains its respective model based on data collected from data producer network functions (NFs). Data producer NFs may be any network function that produces data. For example, one data producer NF may be an access and mobility function (AMF), producing location information in respect of UEs. Alternatively or additionally, MTLFs may train their respective models based on historical data received from a data collection coordination function (DCCF) or from an analytics data repository function (ADRF).

[0060] The AnLF then receives the trained ML models (and/or data indicative of the trained ML models) for a specific analytic ID from the MTLF corresponding to that analytic ID. The AnLF uses the model, potentially in combination with new data from the data producer NF(s) and/or DCCF, to derive analytics that have been requested by a consumer NF.

[0061] The AnLF may subscribe to the MTLF to receive trained ML models for an Analytic ID using a specific API (prior art). The AnLF subscribes based on an analytics

request from a consumer NF. An NWDAF may support both Analytics inference and ML Model Training functions.

[0062] In some cases it may not be possible to exchange data directly between data producer NFs and the NWDAF (i.e. MTLF or AnLF). This may for example be because of privacy concerns: an operator of an edge network or a private slice may not want to share data to external networks. As another example, this may be because of a high signalling load: signalling load can be increased considerably when data is provided to a centralized training function such as that provided by the MTLFs. It can be more efficient for a network at the edge to receive trained models rather than collect data to derive analytics.

[0063] One way for addressing situations in which data is not to be directly exchanged (for example because of privacy concerns or signalling load efficiency) is via federated learning. Federated learning allows local model training functions to exchange model parameters and aggregate trained model and/or intermediate training results, instead of centrally training models by collecting raw data. Federated learning is thus a distributed machine learning framework that allows a model to be trained collectively from data that is distributed across different data owners. This provides an advantage that AI/ML models can be trained closer to the data source, rather than sending raw data to a centralised training node. Instead of sending raw data, parameters, weights or intermediate results of a ML model can be sent back to the centralized node to assist generic model training.

[0064] It will be understood that federated learning can be conceptually divided into horizontal and vertical federated learning. These are schematically depicted in Figure 4A and 4B, respectively.

[0065] Horizontal federated learning, or sample-based federated learning, can be applied in scenarios having data sets that share the same feature space (e.g. each data set includes the same features) but have different samples (for example being collected from different users). This is illustrated in Figure 4A.

[0066] Vertical federated learning, or feature-based federated learning, can be applied in scenarios having data sets that share the same sample space (e.g. each applies to the same

set of users) but differ in feature space (e.g. different features are collected in each data set). This is illustrated in Figure 4B.

[0067] In both horizontal and vertical federated learning models, parameters from each local model training function are sent to a model aggregator to calculate an aggregate model. The model aggregator provides updated model parameters to each MTLF that each MTLF uses to re-train its own model thus allowing every local model training function to have a trained model using data from multiple sources.

[0068] There are different approaches of how to train a model with vertical federated learning (VFL), including “split” and “non-split”. Examples of these approaches are now described in Vertical Federated Learning: Taxonomies, Threats, and Prospects, Qun Li et.al. In particular, Figure 2 thereof depicts a taxonomy of VFL algorithms including split and non-split. Figure 3 thereof depicts non-split VFL in more detail, Figure 4 thereof depicts split VFL.

[0069] In both split and non-split VFL, one party is the “label owner” or “active participant”. This party has the ability to classify input data. The other parties are “passive participants” or “workers” which participate in the VFL training process.

[0070] In non-split VFL, passive participants send intermediate results based on data collected locally using their own model to active participants and the active participants computes gradients/losses using as basis the labels and its own ML model. Some scenarios also include a coordinator that ensures exchanges of messages between VFL parties are encrypted (as depicted in Li Fig. 3).

[0071] In split VFL, in contrast, the model is split between several parties. One party own the top model (label owner/VLF server) and other parties own one or more bottom models (passive participants). This is illustrated in Li Fig. 4. The label owner may also be the active participant. In contrast to non-split VFL, the VFL server is aware of the labels and is able to compute gradient/losses that are shared to passive participants.

[0072] Figures 5A and 5B depict communication flows for federated learning amongst different NWDAFs. These communication flows are described in 3GPP TS 23.288.

[0073] Figure 5A depicts the registration, discovery and selection of clients for federated learning. Each MTLF registers its FL capability (FL client or FL server) within the NWDAF profile in the NRF. The profile includes additionally information on the analytic IDs supported for FL and/or location area and/or Event ID(s) where data collection is supported. The NWDAF MTLF supporting FL server capability discovers candidate FL clients for training a model using federated learning by receiving a list of FL clients from the NRF that support the analytic ID and/or location area and/or Event IDs needed to train the model. The FL server sends a preparation request to all candidates to check if they are capable to join the FL process, i.e., in terms of the data availability and required time schedule, (steps 7-9) and each FL client responds if it can join the process or not.

[0074] Fig. 5B depicts the procedures for FL once the FL clients have been selected and joined the FL process.

[0075] The general procedure for FL is that the FL server sends a subscription request related to federated learning for model training to each selected FL client including local model parameters and each FL client trains the model using data from local NFs.

[0076] Examples of the present disclosure provide solutions to support vertical federated learning enabling the 5GS to assist in collaborative AI/ML operation involving the 5GC/NWDAF or AF. In particular, in order to support VFL, the present disclosure provides methods for VFL entities (either server, active participants or passive participants) to participate in the VFL training process. The following description also sets out actions that the server, active participant may take after a VFL training process is completed (e.g. model information storage).

[0077] Figure 6 depicts various architectures for performing VFL according to examples of the present disclosure. These architectures depict communication between a VFL server 610, an active participant 615 and passive participants 625. In this example, a VFL server 710, which may also be termed a VFL coordinator or controller, is a function that manages the VFL procedures, selects VFL participants 615, 625 for model training, and assigns functions to act as VFL active participant 615 or passive participant 625.

[0078] In this example, a VFL active participant 615 is a VFL function that owns part of a ML model for a given analytic ID and knows the labels for the ML model. The active participant 615 is the main function for training an ML model for a given analytic ID.

[0079] In this example, a VFL passive participant 625 is a VFL function that owns part of an ML model for a given analytic ID, does not know the labels of the ML model, but is able to collect local data for one or more features.

[0080] A given VFL function in the 3GPP network may support a combination of the functions above, e.g., support both VFL server and active participant functions.

[0081] Some examples of the present disclosure relate to scenarios in which once a model is trained using VFL, the VFL server 610 or active participant 615 is aware of all VFL participants (active 615 and passive 625) of the model training process and the contribution weights of each participant.

[0082] Aspects of such examples relate to actions performed in response to a consumer (e.g. an AnLF 605) requesting inference results. Examples of such aspects will now be described with reference to Figures 7A and 7B.

[0083] In a first example, depicted in Figure 7A, the consumer 605 sends a request to the VFL server/active participant 610, 615 for an inference result based on the knowledge that the ML model was trained using VFL. The VFL server/active participant 610, 615 then sends requests for an inference value from each participant of the VFL training process based on the knowledge of the participants in the VFL process when the model was trained.

[0084] In a second example, depicted in Figure 7B, the consumer 605 directly sends requests for an inference value from each participant of the VFL training process based on the knowledge that the ML model was trained using VFL and based on the knowledge of the participants in the VFL process when the model was trained.

[0085] Depending on which of these examples is followed, the consumer 605 or the VFL server/active participant 610, 615 derives an aggregate inference value based on the knowledge of the contribution weights from each participant of the VFL model training process.

[0086] In these examples, a given VFL function (i.e. a VFL server 610, active participant 615, and/or passive participant 625) may be part of the NWDAF AnLF, NWDAF MTLF, or a function in an AF. When interaction between VFL functions from NWDAF and VFL function from AF takes place, then an exchange of messages may be performed via a network exposure function (NEF) 628.

[0087] Figure 6 sets out four architecture scenarios including a VFL server 610, active participants 615, passive participants 625, with involvement of NWDAF AnLF, MTLF and AF. Specifically, the top row of Figure 6 depicts examples in which VFL is performed within a core network, and the bottom row depicts examples in which VFL is initiated outside the core network. Within those rows, the left hand side of Figure 6 depicts examples in which a VFL server 610 is included, and the right hand side depicts examples in which no VFL server 610 is used.

[0088] Figure 7A and 7B depict process flows in accordance with aspects of the present disclosure. The process flows may implement aspects of the wireless communications system 100 as described with reference to Figure 1, and more particularly may be implemented within one or more of the architectures depicted in Figure 6. The process flows may involve one or more of a NWDAF AnLF 705, a VFL server 710, a VFL active participant 715, a VFL passive participant 725, a VFL passive participant 735, a NRF 745 and a ADRF 750. The process flow can thus be implemented within the architecture depicted in Figure 6, with same-named entities having the same functions as described above. In the following description, the operations performed by one or more of the entities 705-750 may be performed in a different order from that shown, or at different times. Some operations may also be omitted from the process flow, and other operations may be added.

[0089] In the examples of Figure 7A and 7B, the contribution to a VFL process of active and passive participant functions 715, 725, 735 can be determined, and contribution weights ascertained for each participant. These can be provided to a model consumer, thereby allowing the consumer to use the model (that was trained using VLF) for inference. The process flows of Figures 7A and 7B thus provide effective ways for performing inference based on VFL-trained models in a wireless communication network. As

explained above, this improves the accuracy and performance of network analytics, thereby improving system performance.

[0090] A first method of inference will now be described with reference to Figure 7A.

[0091] At 1, an NWDAF AnLF 705 supporting AnLF receives a request for analytics (e.g. Observed Service Experience Analytics) and includes analytics ID and analytics filters as specified in 3GPP TS 23.288.

[0092] At 2, the NWDAF AnLF 705 identifies the ML model needed to derive analytics and determines distributed inference is needed (or alternatively determines that the ML model was trained using federated learning). The NWDAF AnLF 705 obtains this information by interfacing with the NRF 745 or ADRF 750 or via previous interaction with a MTLF acting as a VFL server 710 or VFL active participant 715 that has provided the NWDAF AnLF 705 with information of the ML model details. The ML model details include address of the VFL server 710 or the VFL active participant 715 or passive participants 725, 735, contribution weights, and VFL coordination ID (which is used by each VFL participant to identify the specific VFL training collaborative process for training an ML model for an analytic ID).

[0093] At 3, the NWDAF AnLF 705 sends an Inference Request to the VFL server 710 (if available) or VFL active participant 715. The NWDAF AnLF 705 includes the Analytic ID (e.g. Observed Service Experience), analytics filters (e.g. target UEs, service area, slice information, application information etc) and VFL collaboration ID (if known). The Inference Request may be a subscription request (i.e. provide feedback periodically) or a one-time request.

[0094] At 4, it is assumed that a ML model is already continuously trained and updated using VFL.

[0095] At 5, the VFL server 710 or VFL active participant 715 identifies the ML model linked to the Analytic ID and identifies the VFL participants (active 715 and passive 725, 735) that are involved in the vertical federated process for training the ML model and the associated VFL collaboration ID. The VFL server 710 or VFL active participant 715

obtains this information locally or via NRF 745 or via obtaining model information from the ADRF 750.

[0096] At 6, the VFL server 710 or VFL active participant 715 sends an Inference Request to each participant of the VFL process. The request may include Analytics ID, filters such as the samples to do the inference calculation, and VFL collaboration ID. The Inference Request may be a subscription request (i.e. provide feedback periodically) or a one-time request.

[0097] At 7, each VFL participant 715, 725, 735 identifies the local ML model linked to the ML training process using VFL and computes an inference output using its local ground truth data (7a, 7b, 7c).

[0098] In some examples, if the VFL server 710 is not deployed, the VFL active participant 715 does not send an Inference Request but computes locally an inference output.

[0099] At 8, the inference output (or intermediate result) from each participant 715, 725, 735 is sent to the VFL server 710 and/or VFL active participant 715.

[0100] At 9, the VFL server 710 and/or active participant 715 compute an aggregate inference output taking into account the contribution weights from each participant 715, 725, 735.

[0101] At 10, the VFL server 710 and/or active participant 715 may also use the inference output as input to further train the ML model and provides gradient/losses to each VFL participant 715, 725, 735 that each VFL participant 715, 725, 735 can use to further train their local ML model.

[0102] At 11, the VFL server 710 and/or active participant 715 prepares an inference response, including output data, to transmit to the NWDAF AnLF 705.

[0103] At 12, the NWDAF AnLF 705 prepares analytic output data taking into account inference information from the VFL server/active participant 710, 715.

[0104] At 13, the result is sent to the analytics consumer.

[0105] The method of Figure 7A thus provides an effective way for providing analytics using inference from a model that was trained using VFL.

[0106] Figure 7B depicts an alternative example in which the NWDAF AnLF 705 may act as a VFL server 710. In this case the NWDAF AnLF 705 may directly send an inference request to each VFL participant 715, 725, 735. This will now be described.

[0107] At 1, an NWDAF 705 supporting AnLF receives a request for analytics (e.g. Observed Service Experience Analytics) and includes analytics ID and analytics filters as specified in 3GPP TS 23.288.

[0108] At 2, the NWDAF AnLF 705 identifies the ML model linked to the Analytic ID to derive analytics and determines distributed inference is needed (or alternatively determines that the ML model was trained using federated learning). The NWDAF AnLF 705 obtains this information by interfacing with the NRF 745 or ADRF 750 or via previous interaction with a MTLF acting as a VFL server 710 or VFL active participant 715 that has provided the NWDAF AnLF 705 with information of the ML model details. The ML model details include address of VFL server/active participants 710, 715 or passive participants 725, 735, contribution weights, VFL coordination ID (which is used by each VFL participant to identify the specific VFL training collaborative process for training an ML model for an analytic ID).

[0109] At 3, it is assumed that a ML model is already continuously trained and updated using VFL.

[0110] At 4, the NWDAF AnLF 705 sends an Inference Request to each participant 715, 725, 735 of the VFL process. The request may include Analytics ID, filters such as the samples to do the inference calculation, VFL collaboration ID. The Inference Request may be a subscription request (i.e. provide feedback periodically) or a one-time request.

[0111] At 5, each VFL participant 715, 725, 735 identifies the local ML model linked to the ML training process using VFL and computes an inference output using its local ground truth data (5a, 5b, 5c)

[0112] In an example, if the VFL server 710 is not deployed, the VFL active participant 715 does not send an Inference Request but computes locally an inference output.

[0113] At 6, the inference output (or intermediate result) from each participant is sent to the VFL server 710 and/or VFL active participant 715.

[0114] At 7, the NWDAF AnLF 705 acting as a VFL server 710 computes an aggregate inference output taking into account the contribution weights from each participant and prepares Analytic Output data taking into account inference information from the VFL server/active participant 710, 715.

[0115] At 8, the result is sent to the analytics consumer.

[0116] The method of Figure 7B thus provides an effective way for providing analytics using inference from a model that was trained using VFL.

[0117] Figure 8 illustrates an example of a NE 800 in accordance with aspects of the present disclosure. The NE 800 may include a processor 802, a memory 804, a controller 806, and a transceiver 808. The processor 802, the memory 804, the controller 806, or the transceiver 808, or various combinations thereof or various components thereof may be examples of means for performing various aspects of the present disclosure as described herein. These components may be coupled (e.g., operatively, communicatively, functionally, electronically, electrically) via one or more interfaces.

[0118] The processor 802, the memory 804, the controller 806, or the transceiver 808, or various combinations or components thereof may be implemented in hardware (e.g., circuitry). The hardware may include a processor, a digital signal processor (DSP), an application-specific integrated circuit (ASIC), or other programmable logic device, or any combination thereof configured as or otherwise supporting a means for performing the functions described in the present disclosure.

[0119] The processor 802 may include an intelligent hardware device (e.g., a general-purpose processor, a DSP, a CPU, an ASIC, an FPGA, or any combination thereof). In some implementations, the processor 802 may be configured to operate the memory 804. In some other implementations, the memory 804 may be integrated into the processor 802.

The processor 802 may be configured to execute computer-readable instructions stored in the memory 804 to cause the NE 800 to perform various functions of the present disclosure.

[0120] The memory 804 may include volatile or non-volatile memory. The memory 804 may store computer-readable, computer-executable code including instructions when executed by the processor 802 cause the NE 800 to perform various functions described herein. The code may be stored in a non-transitory computer-readable medium such as the memory 804 or another type of memory. Computer-readable media includes both non-transitory computer storage media and communication media including any medium that facilitates transfer of a computer program from one place to another. A non-transitory storage medium may be any available medium that may be accessed by a general-purpose or special-purpose computer.

[0121] In some implementations, the processor 802 and the memory 804 coupled with the processor 802 may be configured to cause the NE 800 to perform one or more of the functions described herein (e.g., executing, by the processor 802, instructions stored in the memory 804). For example, the processor 802 may support wireless communication at the NE 800 in accordance with examples as disclosed herein. The NE 800 may be configured to support a means for obtaining a request message to provide an inference output for a network analytics operation identified by an analytic identifier; outputting an inference request message to each of a set of one or more model training functions, set model training functions being participants in a federated learning training process for a trained model corresponding to the network analytics operation; obtaining an inference output from each function of said set; outputting a response message comprising an aggregate inference output, said aggregate inference output being based on a respective contribution weight, in the federated learning training process, for each said function.

[0122] The controller 806 may manage input and output signals for the NE 800. The controller 806 may also manage peripherals not integrated into the NE 800. In some implementations, the controller 806 may utilize an operating system such as iOS®, ANDROID®, WINDOWS®, or other operating systems. In some implementations, the controller 806 may be implemented as part of the processor 802.

[0123] In some implementations, the NE 800 may include at least one transceiver 808. In some other implementations, the NE 800 may have more than one transceiver 808. The transceiver 808 may represent a wireless transceiver. The transceiver 808 may include one or more receiver chains 810, one or more transmitter chains 812, or a combination thereof.

[0124] A receiver chain 810 may be configured to receive signals (e.g., control information, data, packets) over a wireless medium. For example, the receiver chain 810 may include one or more antennas for receive the signal over the air or wireless medium. The receiver chain 810 may include at least one amplifier (e.g., a low-noise amplifier (LNA)) configured to amplify the received signal. The receiver chain 810 may include at least one demodulator configured to demodulate the received signal and obtain the transmitted data by reversing the modulation technique applied during transmission of the signal. The receiver chain 810 may include at least one decoder for decoding and/or processing the demodulated signal to receive the transmitted data.

[0125] A transmitter chain 812 may be configured to generate and transmit signals (e.g., control information, data, packets). The transmitter chain 812 may include at least one modulator for modulating data onto a carrier signal, preparing the signal for transmission over a wireless medium. The at least one modulator may be configured to support one or more techniques such as amplitude modulation (AM), frequency modulation (FM), or digital modulation schemes like phase-shift keying (PSK) or quadrature amplitude modulation (QAM). The transmitter chain 812 may also include at least one power amplifier configured to amplify the modulated signal to an appropriate power level suitable for transmission over the wireless medium. The transmitter chain 812 may also include one or more antennas for transmitting the amplified signal into the air or wireless medium.

[0126] Figure 9 illustrates a flowchart of a method in accordance with aspects of the present disclosure. The operations of the method may be implemented by a NE as described herein. In some implementations, the NE may execute a set of instructions to control the function elements of the NE to perform the described functions. The operations of the method may be performed in accordance with examples as described herein. In some implementations, aspects of the operations may be performed by a NE as described with

reference to Figure 8. One or more of the operations may be omitted, or performed in a different order from the example shown.

[0127] At 902, the method may include obtaining a request message to provide an inference output for a network analytics operation identified by an analytic identifier;

[0128] At 904, the method may include outputting an inference request message to each of a set of one or more model training functions, set model training functions being participants in a federated learning training process for a trained model corresponding to the network analytics operation;

[0129] At 906, the method may include obtaining an inference output from each function of said set;

[0130] At 908, the method may include outputting a response message comprising an aggregate inference output, said aggregate inference output being based on a respective contribution weight, in the federated learning training process, for each said function.

[0131] It should be noted that the method described herein describes a possible implementation, and that the operations and the steps may be rearranged or otherwise modified and that other implementations are possible.

[0132] Figure 10 schematically depicts a wireless communication system 1005 according to aspects of the present disclosure. The wireless communication system 1005 comprises a first network entity 1010 and a second network entity 1015.

[0133] The first network entity 1010 is configured to obtain a first request message to provide an inference output for a network analytics operation identified by an analytic identifier; output a second request message to the second network entity to perform an inference operation in respect of the network analytics operation.

[0134] The second network entity 1015 is configured to: output a third request message to each of one or more functions to participate in a federated learning training process for a model to perform said network analytics operation; obtain an inference output from each of said one or more functions; output to the first network entity an aggregated inference output

based on said obtained inference outputs and respective contribution weights, in the federated learning training process, for each said function.

[0135] The first network entity 1010 is configured to output a response message comprising said aggregate inference output.

[0136] The description herein is provided to enable a person having ordinary skill in the art to make or use the disclosure. Various modifications to the disclosure will be apparent to a person having ordinary skill in the art, and the generic principles defined herein may be applied to other variations without departing from the scope of the disclosure. Thus, the disclosure is not limited to the examples and designs described herein but is to be accorded the broadest scope consistent with the principles and novel features disclosed herein.