

Sustainable Data Management practices in Federated Learning

Uma Perumal, Vasantharajan Renganathan

Sri Venkateswara College of Engineering, Chennai, India.

Corresponding Author

Email Id: - rajavasanth2079@rediffmail.com

ABSTRACT

The Challenges faced by IoT applications using Federative Learning are dealt with advanced AI algorithms that can work with a combination of logics and various computations among heterogenous local models. The Sustainable data management methodology helps to provide a solution by creating learning models for dealing these challenges as the edge devices available are of with limited resources and capability, but more competence is expected from them. Devices equipped with advanced AI learning methods and practicing Sustainable data management practices can perform better and be a part of Sustainable development goal of a nation.

Keywords:-Federative Learning, Sustainable Data Management, Plugin Module

INTRODUCTION

Rapid development in internet with diversified Internet of Things (IoT) can generate enormous amounts of data, cannot be used for centralized cloud computing as high volume of data communication burdens the network to integrate data from numerous IoT devices. IoT devices capable of Edge computing can bring down the bandwidth required for communication by minimizing the frequency of communication and volume of data by various mechanisms dictated by the AI embedded along with the IoT sensors. Raw data collected by IoT is transmitted to others, leaving a possibility of exposing personal information that is privacy sensitive.

Machine learning (ML) in a large scale with lots of data for analysis from distributed datasets generated by the interconnection of heterogenous devices are with challenges in the area of privacy, communication, biased database, computation costs involved and unbalanced datasets. These obstacles are dealt with advanced machine learning concept like Federated Learning (FL), a distributed, decentralized ML that has no

direct exposure of the individual's private data, usually encrypted for safety and privacy and can use this anonymous data only for a specific purpose of ML, then exposing the individual.

Google in 2016 pioneered in developing a decentralized framework for FL to ease the drawback in centralized or cloud-based data analysis of leakage of client's personal data as they are shared among various organizations resulting in the violation of data privacy agreements, thereby bringing down the client's confidence resulting in the distrust over data networks. The IoT devices are not upgraded, and they simply perform their work of generating raw data, which is not going to leave the device, but the updates of the model are sent to the cloud server, eliminating risk of personal data exposure.

Federated learning is an emerging and trending technology with many domains ranging from Multitask with Meta learning, Generative Adversarial Learning, Reinforcement Learning, Unsupervised Learning, Semi-supervised Learning, Transfer Learning and Federated Averaging.

Single IoT's cannot deliver the output data because of the performance constraints as expected, but a group of one such can be trained collectively to deliver a high-quality model, which empowers every individual participant of the group to learn and benefit from the model generated without any compromise to the privacy of the data privacy. Edge computing devices use one such combination of IoT devices.

Heterogenous datasets collected from various backgrounds like social media, healthcare, legal, business, government, career or work platforms by the organizations face a major challenge in data management in securing data, use the available data for sustained information practice to reduce the data and resources allocated to collect the same and to look for alternate methods to generate a revenue stream for their business from the available data.

TECHNOLOGY

Federated Learning (FL)

Federated Learning is a type of decentralized machine learning in that works on edge devices, whereby the algorithm is trained within the edge servers or devices holding the local raw data without any data exchange with the cloud or central server, but sharing models got out of several training iterations by the algorithm available or send to the edge devices, that gets updated frequently. The models trained locally with several iterations are then sent as an updated version to the central server and retain the dataset within the edge device or infrastructure for the reasons of security and privacy. The role of central server is to integrate all the contributions of the participant edge servers, update the same and again share them back to all the servers, thereby sharing the updated model among various contributors. Incase if one of the contributors quit, the training continue with the remaining others without any major impact as the network is not much reliant on

the data from a single contributor, and the same goes for new entrants, of those who want to join the network. The main aim of FL is to enable every participant to gain global knowledge through the locally shared model.

IoT

IoT is a technology that uses single or multiple sensors that are implanted in a device in addition to software and communication technologies for the intention of sharing/sending information and data with the working methodology of exchanging data with the cloud or server using internet or a wider connected network of computers and communication devices for interaction. These devices can do self-reporting in real time without the need for any human intervention. They work with an active network connection and can transmit data, the network can either be the internet or a WAN/LAN for communication purpose.

Edge Computing

Edge computing is a distributed computing methodology that uses the edge devices enabled with IoT's and embedded processors, where the computing is done in the edge of the network (embedded processors process the data and deliver outputs to cloud servers), which brings down the load of cloud servers in processing data. Raw data is not sent to the cloud servers but processed data mostly for storage purpose as memory available with edge processors are relatively small to the cloud servers.

Convergence

It is defined as a stable point at which there is no further change in the solution obtained from a sequence of optimization in an iterative process. This is the tipping point of learning in the training model, where the rate of learning comes down as the difference between the solution obtained

and the solution obtained in the previous iteration has minimal difference. Any further iteration results in no significant change in the result obtained.

Convergence sometimes don't happen in AI algorithms due to many reasons. For instance, if the network architecture is modified for the sake of expecting a better data mode, then convergence fails. If the nodes available are not enough to transform the input into accurate output, then convergence fail. If the training data collected is inadequate, or if data integrity is compromised during data collection, convergence fails.

Federated averaging (FedAvg)

Federated averaging is a communication algorithm that shares updated parameters among the participating clients using simple weighted average of the updates from a client. The client devices run iterations in Stochastic Gradient Descent (SGD) with the locally available data and compute for an update. This update among all the servers is finalized by the simple weighed average and is called the final update, communicated to the clients. In a distributed training an effective algorithm is used to communicate with many devices, that enables privacy protection by keeping the data locally and establishing a central parameter server that communicates with all the participants.

Data Lifecycle Management (DLCM)

Data gathered and stored in edge/local devices have a use for a certain period and after which, unused or unassessed data reach its end and is removed. The process of transfer of data through its entire lifecycle and securely deleting is called Data Lifecycle Management.

- **Data Capture:** This is the start of the life of a data from IoT sensors or data generated in the utility by the user. It could be an image, video or document and the information about the same is entered into database and is retrieved for specific roles by the client.

- **Backup and Recovery:** In this stage the data is archived in a way it is easy to access with complete accuracy and a thorough backup process to make a copy is implemented as a disaster recovery measure of the available data.

- **Data Management:** This stage of data lifecycle management ensures about the availability of data for generating models or share among the permitted users.

- **Data Retention and Removal:** This is the final phase of data lifecycle management as the data available is discriminated based on the various metrics of quality, usability and each piece of data using backup and recovery process adhering to the compliance and standards. Then the data is archived or destroyed based on the former recommendations.

METHODOLOGY

Data collected by various edge devices consume resources like time and energy to do the work. The data collected must be useful and is made available in such a way that it has optimum data life cycle. Data generated by the IoT devices, applications in the edge devices are stored locally, and a data model is communicated with the cloud server for updating with the global

FL, a method of ML faces a lot of challenges in the areas of Data collection, Choice of data source, Data security and privacy, Data preparation methods, large volume data management, Discovery of data, Insights and Data Lineage with a lot of limitations.

FL methods work well when the volume of data is less or medium, but when the volume is too big, it calls for an update on the hardware or replacement and communication speed as the available resources are loaded too much. The local data over a period becomes bulkier and they are stored as backup in the local devices or any secured cloud storage. This consumes a lot of resources of time and network bandwidth when the size of the file increases abruptly.

The AI algorithms send from the server as updates usually update the local model with the available global from the cloud server. The AI algorithms known as updates or patches are sent to the local devices as updates are really a new application program that can run with the available data after the installation. Recent studies show about the difficulties faced by the clients over the frequent updates with bugs and minor fixations. If the version is unstable, then roll back is done and in due course the data with the app goes on increasing with a significant size and allocation in the edge device memory. Until a stable version is available for a period, unstable updates disrupt the normal working methodology of the clients and if the application becomes unstable, a uninstall following a reinstall of the previous version is suggested by the app developer.

Frequent upgrade on the hardware resources is making a heavy impact on the environment as the Carbon Footprint for the newly manufactured hardware and the disposed existing hardware are accounted by the Sustainable Data Management practices, along with the power consumed by the edge device to manage heavy data from creating backup, reusing backup for data utilization. The cloud centers working to manage or create the local and global model algorithms are also consuming power, comes under the pressure of optimization of available resources.

The Sustainable Data Management practices call for optimizing data management, which command on identification of critical and non-critical data available with the edge device. Discrimination among both can pave way the effective utilization of data management by keeping the non-critical data in a storage device (either within the edge device or an outside storage device) and the critical data in the working model, that regularly consumes power for updating

or utilization by the client.

APPLICATION-1

When the volume of local data is enormous, significant memory space is occupied in the device memory is experienced. Backup data increases the time to retrieve the same and make it available at the time of utilization. For instance., An Engineer/user working in an organization for a period (about 10 years) uses his edge device to aid his work and the device becomes customized according to his work and his searches on the internet are based on his previous preferences and the ML algorithm adapts itself towards the user and provides tailored results according to his search requirements. The same continues even when the edge devices are replaced, as the data grows on size and as a measure of upgradation of resources by the client.

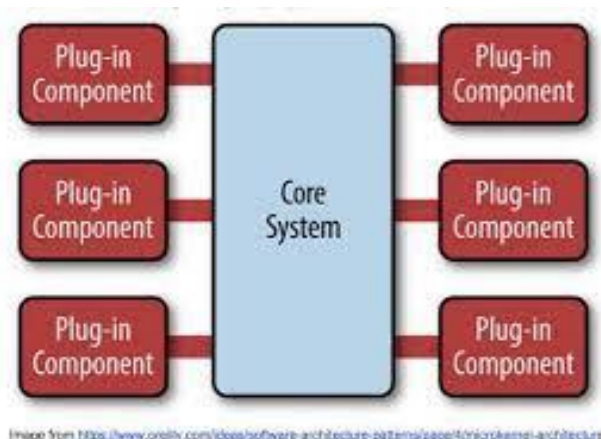
After some time, the Engineer/user is promoted as a manager/user in the organization, and he uses the edge device as per his requirement for his new role in the organization. The device now adapts itself to the new search, preferences and provides results based on his new requirement. The old data (when used as an engineer/user) becomes obsolete over a period as it was not required by the client. The obsolete data is usually discarded and safely deleted in the process of data life cycle and by the available algorithms, as data of no utilization, locally available model, global model are not updated by the FedAvg method as convergence of data occurs in the local device as no new search of relevant topics of the data is required by the client. Hardware upgradation aids in updating the current model, leaving no updating in the old model.

The Manager/user after a period (5 years) is further elevated as a technical consultant/user by the organization and here his role in the organization demands his engineering expertise in addition to his

managerial capability. The user when using the new updated model tries to bring the old, tailored model he used when he was an engineer (looking for more technical related data) finds difficulty in looking for tailored results, he had before a period as the local models updated in his edge device has no or less trace of data of the former role. The user is forced to start from the first, and it takes time and resources for the edge device to train itself to comeback as on with technical device. Duration taken to learn by the FL model has an impact in the service served to the user by the AI based on performance as most of the old data, algorithm model is safely deleted or overwritten by the new

model. Updates or patches of the app can be substituted by Plugin/architecture methodology for effective ML in FL.

Plugin architecture uses collaborative software environment where the application can be created by the combination of different, transformable components that does not rely on each other but can perform when assembled. The architecture consists of a core system and plugin modules. The Core systems consist of its own ecosystem from logic to data access and management, security, logging and utility functionalities. The core is not customized, or no specific customization is implemented in the core.



The Plugins are stand-alone, independent components with specialized processing, additional features and custom implementations that are meant to improve or extend the core capabilities with an ease in flexibility and isolation of application features. The plugins can be removed any time or modified based on the requirement of the user. The core system must know about the plugin modules and the way of exchanging data among plugins. The core system declares extension points and a well-defined interface for interaction. The plugin, core interaction happens by exchange of data like name, input/output handlers etc., The connection between the two is application specific.

Edge devices when upgraded or when the user changes his preference of model (based on the data) over another can use the plugin architecture to make the existing old model a plugin module (PM), and this PM is stored in the external storage device (residing in edge server) or in the edge server with naming depending on the date (e.g., PM1992022 for the module created from the updated module available on 19-9-2022. Ease of renaming is also available i.e., PM with Engineering update for easy recognition by the edge user). PM's can also use data encryption for security and privacy in addition to a compressed format to reduce storage space. Provisions to copy or move the file to any remote storage location (that

does not regularly use power- a Non-volatile memory/NVM to keep the data- thereby reducing energy required to maintain one) that is defined by the user. Many such PM's can be created based on the classification of the user for easy search and identification. Moving the PMs to a Public cloud is a better option than other cloud server, as these public servers run on renewable energy, incorporate new technologies (green energy) and smart methods in infrastructure utilization. For e.g., Cloud Footprint calculator released by Microsoft in one such that can give an idea to the user of choosing the best cloud service.

The app currently available with the user with the latest updated data model is used as a core, as it consists of latest application installed. The user requiring the PM (with the data model) can use the plugin architecture to connect the required data model with the core app, which then transfers the data model into itself and becomes the tailored model required by the user. The user can also connect several PM's until restricted by the core app based on the hardware and software capability of the edge device. The user can also remove the plugin when not needed and the data model in the edge device rolls back to the previous model that was before the plugin.

Plugin architecture makes the existing device efficient even when there is an additional number of data to be added in the data model. This reduces the need for frequent updates required by the app in the system. A PM from any user (Expert in a specific field of study), that has got convergence in its data model can be used as a standard model and used for creating a global model can be kept in cloud server and labelled under classification (using the specific field of study) based on the keyword search by other clients requiring models from the same field. This classification will help ML algorithms to deliver services to

like minded users, aiding them the ease of finding data they look for on the internet.

The data models created using the PMs are available with the user and less effort, resources are used to switch from various PM's as they are readily available, and the core app AI is ready to learn in a short span of time. The data in the data model here is used even after the data life cycle management as the old data model can be used to guide new users of those, may require, thereby reducing the effort spend in developing the new model ensuring a sustainable approach in data management.

APPLICATION-2

Sustainable data management (SDM) practices are followed only in organizations as they are under the supervision of agencies implementing audits in energy, carbon footprint etc., but not by the edge device user. The number of edge device users is so high, that surpasses the entire capacity of the data centers. If the edge user can follow sustainable data management practice, then SDM can take a clear shape, but edge users don't have necessary data, guidelines to do the same. Teaching them to practice is cumbersome, and an AI algorithm that does SDM is the need for the hour.

There are inferences in the web that has been already analyzed by experts about the energy consumption/carbon footprint and these data can be used as a measuring standard by the AI. The AI then regularly updates about various inferences and creates a global model. This model has the capacity to include the following:

- Lean Principle (used in manufacturing engineering) of eliminating data waste or useless data, continuous improvement and can be called Lean Data Management (LDM) as it starts from the very first method starting with working on the right questions before collecting data. LDM emphasizes in

collecting the relevant data for the application, in spite of collecting and sorting data later, as inefficient methods simply consume resources and time. Effective planning and work out is emphasized in every single step as done in manufacturing engineering (because Lean manufacturing includes the cost involved in every step as cost or resource utilization is a prime factor in business). LDM prioritizes in removing waste and minimizing variation of data, a data to effectively deliver the insight than simply filling the space.

- Proper DLCM that can find the relevance of the data collected, trains with lesser data, then archive the same in a remote location or NVM and safely delete irrelevant data. For e.g., Google's new carbon-intelligent platform uses its advanced AI to work with non-critical computing tasks during the time, where there is much energy generated from renewable source is available in the grid, thereby reducing the carbon footprint of data analysis, and this is a measure of effective SDM of data.
- Creating Small datasets from the existing database using data augmentation that can be adjusted to create more useful inputs for training AI models. These datasets can come up with granular insights that can be more useful in decision-making process in an individual level.

-

CONCLUSION

SDM is becoming more energy conscious, environment impact oriented than simple AI to handle data effectively. SDM is now a part of Sustainable Development Goal (SDG) by various nations. Industries in the business of data are given benchmarks to improve their effective use of energy and guidelines by governmental agencies to comply with, like respecting privacy of the user, security to the data available, to bring down carbon footprint, is a protective measure to save our environment, our earth as enough damage was already made by the human in the name of development,

technology. Internet data on carbon footprint per data is alarming and every data generated counts as big data involves in enormous data management and with a multiplication factor of carbon footprint/data with big data may evolve as critical factor in contributing green house gases as most of the countries haven't adopted green energy policy in a full scale.

SDM emphasizes best data management practices to both edge user and cloud servers as they are the stakeholders in FL. As resources are getting scarce and exhausting natural resources are backfiring in the name of climate change. Advanced AI methods imbued with SDM can run the current situation until a new disruptive tech (like Quantum computing) is invented, that can easily deal the limitations of AI and keep the data business running.

REFERENCES

1. https://cs.uwaterloo.ca/~m2nagapp/courses/CS446/1195/Arch_Design_Activity/PlugIn.pdf.
2. Handbook of Sustainability-Driven Business Strategies in Practice <https://doi.org/10.4337/9781789908350>.
3. SUSTAINABLE FEDERATED LEARNING <https://arxiv.org/pdf/2102.11274v1.pdf>
4. Review of Artificial Intelligence and Machine Learning Technologies: Classification, Restrictions, Opportunities and Challenges https://mdpi-res.com/d_attachment/mathematics/mathematics-10-02552/article_deploy/mathematics-10-02552-v2.pdf?version=1659521627
5. https://www.researchgate.net/publication/348977401_Artificial_Intelligence_Learning_and_Limitations.
6. Limitations Of Artificial Intelligence <https://scholarworks.rit.edu/cgi/viewcontent.cgi?article=12113&context=theses>

7. https://www.researchgate.net/publication/317613827_Sustainability_learning_processes_concepts_benchmarking_development_and_integration
8. https://hub-media.aashe.org/uploads/A+Guide+for+Applied+Sustainability+Learning+Projects_v1.0_03.03.17_Final.pdf
9. https://www.academia.edu/8215770/Sustainable_learning_and_ICT_toward_sustainable_development_Theories_and_empirical_studies

AUTHORS PROFILE

Uma Perumal is an Assistant Professor in Computer Science and Engineering. She's got a master's in computer science and Engineering. Field of interest includes Big Data, Data Analytics and Data Science.

Vasantharajan Renganathan is a certified sustainable business agent and a business English professional. He has got a Bachelor's in Engineering and master's in business administration. He is also a freelance Expert and a consultant for online consultancy services.